KU Leuven
Groep Biomedische Wetenschappen
Faculteit Geneeskunde
Centrum Menselijke Erfelijkheid

KARAKTERISERING VAN GENOMISCHE EN TRANSCRIPTOMISCHE LANDSCHAPPEN VAN T-CEL ACUTE LYMFOBLASTISCHE LEUKEMIE DOOR MIDDEL VAN NEXT GENERATION SEQUENCING

Zeynep KALENDER ATAK

Promoter: Prof. Stein AERTS
Co-promoter: Prof. Jan COOLS

Proefschrift voorgedragen tot het behalen van de graad van Doctor in de Biomedische Wetenschappen

December 2013

KU Leuven
Biomedical Sciences Group
Faculty of Medicine
Center for Human Genetics

CHARACTERIZATION OF GENOMIC AND TRANSCRIPTOMIC LANDSCAPES OF T-CELL ACUTE LYMPHOBLASTIC LEUKEMIA WITH NEXT GENERATION SEQUENCING

Zeynep KALENDER ATAK

Jury:

Promoter:	Stein AERTS
Co-promoter:	Jan COOLS
Chair:	Chris MARINE
Secretary:	Diether LAMBRECHTS
Jury members:	Frank CLAESSENS
	Peter VAN LOO
	Jurgen DEL-FAVERO

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Biomedical Sciences

December 2013

ACKNOWLEDGEMENTS

This work would not have been possible without the valuable guidance and contribution of a number of special people.
First and foremost, I would like to thank my supervisor Prof. Stein Aerts. Always supportive and enthusiastic, he made my PhD journey extremely rewarding. Stein showed how 'good science' is done everyday, and I will try my best to live up to the example he set. I would also extend my gratitude to my co-supervisor Prof. Jan Cools. I have learned a lot from our valuable discussions throughout my projects, and his confidence in my work has been a great source of motivation. Apart from my supervisors, I had the opportunity to work with Prof. Kim De Keersmaecker throughout my PhD. Her insights and comments opened new perspectives for me and I am grateful for our invaluable collaboration.

I would like to thank the members of the PhD examining committee and jury, Prof. Chris Marine, Prof. Diether Lambrechts, Prof. Frank Claessens, Prof. Peter Van Loo, Prof. Jurgen Del-Favero, and Prof. Sabine Tejpar. I thank you for the time you devoted to review my manuscript and for your comments and helpful suggestions.

I would like to thank all the former and current members of the LCB: Annelien, Bram, Delphine, Dimitry, George, Gert, Hana, Jelle, Katina, Kris, Lotte, Marina, Mark, Rekin's, and Valerie. Special thanks to Gert, for his technical support through all my projects and Marina, who has been my PhD comrade from the day one. I'm really grateful to have met and worked with all of you.

I would also extend my gratitude to the members of MPL. Especially I would like to thank to Ellen, Valentina, Daphnie, Idoya and Carmen for all our collaborative projects. It has been great to learn from you and great to discover things together. Further, I thank Sonja for the administrative support and assistance in many different ways.

Ve son olarak da aileme tesekkur etmek istiyorum. Annecigim, babacigim ve kardesim; doktora yillarimda uzakta olmaniz hic bir seyi degistirmedi, siz hep benim yanimda oldunuz. Sizin kosulsuz desteginiz ve sevginize sahip oldugum icin minnetarim. Sevgili kocacigim Onur ve minik kizim Arya, her yeni gune hevesle ve mutlulukla baslama sebeplerim, sizlerin varligi ve destegi (Onur'un biraz daha fazla) benim en buyuk motivasyonum oldu, iyi ki varsiniz.

SUMMARY

Cancer is a genetic disease caused by the gradual accumulation of somatic mutations in key driver genes. Next generation sequencing (NGS) technologies enable comprehensive characterization mutational landscapes of tumor genomes, lead to the discovery of cancer drivers and eventually contributing to a better understanding of tumorigenesis.

In this thesis, we have used various NGS technologies to decipher T-cell acute lymphoblastic leukemia (T-ALL) genomes. T-ALL is a type of leukemia originating from malignant transformation of developing T-cells as a result of chromosomal translocations and cooperating point mutations.

We have identified novel T-ALL drivers using three NGS experiments, namely targeted sequencing, exome sequencing and transcriptome sequencing. We assessed available tools and methods and constructed bioinformatics pipelines for each experiment. We further validated and optimized these pipelines to obtain the most accurate and specific predictions sets. We evaluated our findings in the view of existing knowledge and put forward novel driver processes and genes taking part in T-ALL pathogenesis.

In our first project we followed a targeted sequencing approach and sequenced 97 genes across 15 primary samples and 18 T-ALL cell lines. By conducting a benchmark study we identified the most optimal bioinformatics pipeline for this experimental setup and obtained high quality mutation prediction. We have found mutations in the known T-ALL drivers as well as in promising candidates such as *TET1*, *JAK3* and sprouty family members *SPRY3* and *SPRY4*.

In our second project, we have sequenced all coding exons of 67 primary, 39 matched remission samples and 17 cell lines. We have implemented somatic mutation calling methods and rigorous filtering to characterize the mutational profile of these T-ALL *exomes*. We observed marked differences between the adult and pediatric samples both in terms of mutation rate and mutation pattern. Eventually, we identified 15 driver genes affected by somatic mutations, including 7 novel drivers. From this list, we discovered the involvement of the ribosome in tumorigenesis via mutations in *RPL10* and *RPL5*, and also identified a novel tumor-suppressor, *CNOT3*.

In our third project, we focused on the transcriptome and applied RNA-seq technology on 31 primary samples and 18 cell lines. We have assembled, optimized and validated bioinformatics pipelines for the measurement of gene expression levels and the discovery of mutations, fusion genes and alternative transcript events. This approach enabled a complete characterization of the T-ALL transcriptome, revealing novel drivers and events including the oncogenic fusions *SSBP2-FER* and *TPM3-JAK2*; an exon skipping event in *SUZ12* and mutations in candidate T-ALL drivers including *STAT5B*, *PTK2B* and *H3F3A*.

In this thesis, we discovered novel T-ALL drivers, assessed and implemented state of

the art bioinformatics approaches and demonstrated the power of NGS technologies in the context of cancer genomics. Ongoing and future sequencing projects are expected to offer further insights into human cancers and eventually lead to advancements in personalized cancer treatment.

SAMENVATTING

Kanker is een genetische aandoening veroorzaakt door somatische mutaties die zich accumuleren in potentiële oncogenen. Next generation sequencing (NGS) technologieën maken het mogelijk om mutationele landschappen binnenin een kankergenoom te karakteriseren, om genen met oncogene drijfkracht te ontdekken, en om een dieper inzicht te verwerven in het proces van oncogenese.
In deze thesis hebben we gebruik gemaakt van NGS technologieën om het genoom van T-cell acute lymphoblastische leukemie (T-ALL) te ontcijferen. T-ALL is een klasse van leukemie die ontstaat door de kwaadaardige transformatie van ontwikkelende T-cellen ten gevolge van chromosomale translocaties en punt mutaties. Door toepassing van drie specifieke NGS-gebaseerde projecten hebben we nieuwe kandidaat kankergenen geïdentificeerd die het proces van T-ALL kunnen aandrijven. Hierbij hebben we bioinformatica methodes vergeleken en voor elk van de NGS experimenten vervolgens een bioinformatica protocol opgesteld. We hebben vervolgens deze protocols gevalideerd en geoptimaliseerd om zo de meest accurate en specifieke voorspellingen te bekomen. We hebben onze bevindingen vervolgens geëvalueerd in de context van de huidige kennis over T-ALL en schuiven aan de hand daarvan nieuwe kandidaat genen en processen naar voor die een oncogeen aandrijvend potentieel hebben in de context van T-ALL.
In het eerste project hebben we een doelgerichte sequencing-aanpak gehanteerd waarbij we 97 genen in 15 primaire stalen en 18 T-ALL cel lijnen gesequenced hebben. Aan de hand van een referentiestudie hebben we het meest optimale bioinformatica protocol opgesteld passende bij deze experimentele opstelling. Dit leverde ons mutationele voorspellingen op van hoge kwaliteit zoals mutaties in gekende T-ALL oncogenen alsook in veelbelovende nieuwe kandidaat genen zoals *TET1*, *JAK3* en de sprouty familie leden *SPRY* en *SPRY4*.
In het tweede project hebben we het ganse coderende genoom, of exoom, gesequenced van 67 primaire kankerstalen, 39 overeenkomstige stalen in remissie en 17 cellijnen. Hierbij hebben we methodes geimplementeerd die in staat zijn om mutaties te identificeren gevolgd door een zeer nauwgezette filtering om zo het mutationele profiel van de T-ALL exomen te karakteriseren. We hebben opvallende verschillen geobserveerd tussen volwassen en pediatrische stalen, zowel op gebied van de mutatie snelheid en het mutationele patroon. Finaal hebben we een lijst met 15 potentiele oncogenen geïdentificeerd met een somatische mutatie, waarvan 7 nieuwe kandidaten. Zo hebben we mutaties ontdekt in de genen *RPL10* en *RPL5*, wat wijst op een belangrijke rol voor ribosomen in het oncogene proces. Daarnaast hebben we ook een nieuwe potentieel tumor-suppressieve functie blootgelegd voor het gen *CNOT3*.
In het derde project hebben we ons voornamelijk gefocust op het transcriptoom van T-ALL. Hiervoor hebben we de RNA-sequencing technologie gebruikt op 31 primaire

stalen en 18 cellijnen. Ook hier hebben we een analytisch protocol opgesteld, geoptimaliseerd en gevalideerd op basis van de bioinformatica methodes die nodig zijn om genexpressie te meten en om mutaties, fusie-genen en alternatieve transcripten aan het licht te brengen. Met deze aanpak waren we in staat om een volledige karakterisatie te bekomen van het transcriptoom van T-ALL wat uiteindelijk geleid heeft tot de identificatie van nieuwe kandidaat genen met oncogene drijfkracht. Daarnaast hebben we ook nieuwe oncogene fusies zoals *SSBP2-FER* en *TPM3-JAK2*, exon-overslaande events in SUZ12 en mutaties in *STAT5B*, *PTK2B* en *H3F3A* geïdentificeerd.

In deze scriptie hebben we nieuwe kandidaat genen ontdekt met het potentieel om het oncogene proces in T-ALL in gang te zetten. Daarnaast hebben we een grensverleggend bioinformatica protocol opgesteld, beoordeeld en toegepast en daarbij de kracht van NGS technologieen in de context van *kankergenomica* aangetoond. Verwacht wordt dat huidige en toekomstige sequencing-projecten ons alsnog een diepgaander inzicht zullen geven in het humane kankerproces en een immense vooruitgang zullen betekenen in het gevecht tegen kanker.

RELATED PUBLICATIONS

This thesis is based on the following papers, which are referred to text by their Roman numerals.
[#] Authors contributed equally

I. **Zeynep Kalender Atak**[#], Kim De Keersmaecker[#], Valentina Gianfelici, Ellen Geerdens, Roel Vandepoel, Daphnie Pauwels, Michaël Porcu, Idoya Lahortiga, Vanessa Brys, Willy G. Dirks, Hilmar Quentmeier, Jacqueline Cloos, Harry Cuppens, Anne Uyttebroeck, Peter Vandenberghe, Jan Cools, Stein Aerts. High Accuracy Mutation Detection in Leukemia on a Selected Panel of Cancer Genes.
PLoS ONE **7,** e38463 (2012).

II. Kim De Keersmaecker[#], **Zeynep Kalender Atak**[#], Ning Li[#], Carmen Vicente, Stephanie Patchett, Tiziana Girardi, Valentina Gianfelici, Ellen Geerdens, Emmanuelle Clappier, Michaël Porcu, Idoya Lahortiga, Rossella Lucà, Jiekun Yan, Gert Hulselmans, Hilde Vranckx, Roel Vandepoel, Bram Sweron, Kris Jacobs, Nicole Mentens, Iwona Wlodarska, Barbara Cauwelier, Jacqueline Cloos, Jean Soulier, Anne Uyttebroeck, Claudia Bagni, Bassem A Hassan, Peter Vandenberghe, Arlen W Johnson, Stein Aerts & Jan Cools. Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia.
Nat Genet 45**,** 186–190 (2013).

III. **Zeynep Kalender Atak**[#], Valentina Gianfelici[#], Gert Hulselmans[#], Kim De Keersmaecker[#], Arun George Devasia, Ellen Geerdens, Nicole Mentens, Sabina Chiaretti, Kaat Durinck, Anne Uyttebroeck, Peter Vandenberghe, Iwona Wlodarska, Jacqueline Cloos, Robin Foà, Frank Speleman, Jan Cools, and Stein Aerts. Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia.
Accepted for publication in *PLoS Genetics*.

TABLE OF CONTENTS

CHAPTER I: INTRODUCTION

CANCER PROCESSES

Hanahan and Weinberg defined the "hallmarks of cancer" that constitute the essential changes in cell physiology necessary for tumor development and growth [1,2]. These principles are useful for simplifying the dauntingly complex disease of cancer. We can evaluate the NGS findings from the viewpoint of this framework and understand the contribution of NGS in cancer genomics. In the original paper, the authors have defined six hallmarks that are shared essentially by all human cancers: "self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis" [1]. Progress in the cancer research field resulted in the addition of two hallmarks in 2011 "reprogramming of energy metabolism and evasion of immune destruction" [2]. The authors also elaborate on "enabling characteristics" that facilitate cancer cells to acquire hallmark capabilities, namely genome instability and mutation, and tumor promoting inflammation. Collectively these hallmarks dictate tumor formation and proliferation.

Genome instability

The DNA repair and maintenance mechanism of the cell is almost immaculate and allows very low rates of mistakes during each cell division (10^{-9} for somatic cells [3], and 10^{-11} for stem cells [4]), thus tumor cells may need to circumvent these mechanisms to obtain favorable genotypes. Although next generation sequencing studies revealed that some pediatric cancers and liquid tumors might achieve neoplastic transformation with low mutation loads [5], many tumors increase mutation capacity to facilitate the acquisition of hallmark capabilities and promote tumor growth [6]. Increased mutation capacity of the tumors facilitates acquisition of the hallmark capabilities and promotes tumor growth. Genomic instability can be achieved by several mechanisms such as exposure to mutagens, inactivation of the "caretaker" genes [7], which are DNA repair, DNA damage response and mitotic checkpoint genes; as well as oncogene induced DNA damage and telomere disruption [1,2,8,9]. NGS studies confirmed the existence of various types of genomic instability inherent in cancer genomes, and in some cases elucidated the mechanisms governing it.

Chromosomal instability (CIN) is the major form of genomic instability in human cancers [10],consisting of alterations in chromosome numbers (aneuploidy), chromosome translocations, and gene amplifications. One of the contributions of NGS technologies in detection of CIN is the high-resolution it provides. We can now observe chromosomal alterations that were not detectable with previous methods. For instance, NGS analysis revealed a large number of intra-chromosomal rearrangements and tandem duplications in breast cancer genomes, which would not have been detected with low resolution cytogenic methods [11]. Furthermore, a new CIN phenomenon termed "chromothripsis" was identified through NGS [12]. It involves up to thousand chromosomal rearrangements in a single event localized in one or few chromosomes, and it is observed in around 3% of cancer genomes across all subtypes, with a higher prevalence in bone cancers (25%) [13].

Moreover, comparative evaluation of CIN trends across cancers is now possible with the results of NGS studies spanning many human cancers. In the pre-NGS era, epithelial tumors were thought to be more genomically unstable with higher proportion of chromosomal rearrangements compared to the hematological and mesenchymal cancers [14]. Analysis of NGS studies across 18 cancer types confirmed this observation [15]. It was also revealed that CIN profiles could be shared across different cancers and show dramatic changes between individuals of the same cancer type. For instance, BRCA-associated breast and ovarian cancers have genomic instability due to defects in homology directed repair and have high levels of CIN and a similar spectrum of mutations [16], while lung cancer in smokers and never smokers display extensive differences in the number of alterations due to mutagens in tobacco [17].

Another form of genomic instability involves alterations spanning single nucleotides or short fragments, thus termed "mutational instability" [10]. Mutational instability can manifest itself with an abnormal rate of somatic mutation. NGS analysis revealed that melanoma and lung cancer genomes harbor around 200 non-synonymous mutations per tumor, 3-fold higher than in solid tumors [5]. These cancers are associated with ultraviolet light and tobacco exposure, respectively, both of which cause DNA damage with different mechanisms [18,19]. Germline defects in caretaker genes have been associated with hypermutation profile in hereditary [8] and sporadic cancers. For instance in colon cancers, the high mutational load is caused by alterations in the caretaker genes, including hypermethylation of the *MLH1* gene and mutations in the components of DNA mismatch repair pathway (*MLH1, MLH3, MSH2, MSH3, MSH6 and PMS2*) together with mutations in *POLE* (proofreading domain of polymerase) [20]. These tumors harbor a 7-fold higher number of non-synonymous mutations than the mutagen induced lung and melanoma tumors [5]. Furthermore, localized hypermutation patterns, termed "kataegis", were identified and observed in breast cancer with NGS analysis. This phenomenon is attributed to the AID/APOBEC family of proteins, which originally function as RNA editing enzymes (C-to-U) with recent findings proposing a DNA mutator role [21,22].

Tumor promoting inflammation

The presence of immune cells around tumor tissues has been observed in ranging quantities across almost all tumor types [2]. Although the presence of these cells might indicate the efforts of the immune system to fight with tumors, there is growing evidence that the presence of these cells is protumorigenic and contribute to the acquisition of the hallmark capabilities by providing bioactive molecules such as growth or survival factors [23]. Inflammation related programs in the tumor cells are predominantly orchestrated by the activated oncogenes *NFKB* and *STAT3* [24]. In this context, activating mutations in *NFKB* across many cancers [24] and constitutive phosphorylation of *STAT3* downstream of *EGFR* mutations in lung cancers [25] can exemplify genomic aberrations associated with inflammation. In addition to providing a favorable environment for tumor growth, it has been suggested that activated inflammatory cells can increase mutation rates of tumor cells either through the release of DNA damaging agents such as reactive oxygen species (ROS) and reactive nitrogen intermediates (RIN); or by stimulating ROS accumulation in the neighboring epithelial cells [26]. In fact, mutation rates are found to be 4-fold higher in the inflamed microenvironment compared to the normal tissues [27], thus linking these two enabling characteristics of cancer together.

Self-sufficiency in growth signals

Self-sufficiency in growth signaling was the first hallmark defined by the authors. Normal cells are tightly regulated by external signaling mechanisms for proliferation and growth. Cancer cells circumvent the necessity of these external signals by producing their own growth factors, stimulating nearby cells to supply growth factors, increasing the number of receptor proteins in the cell surface to become hypersensitive to ambient levels of growth factors, and finally by deregulating signaling pathways to be constitutively active without the presence of ligand [2].

NGS studies provided ample evidence in regards to the deregulation of growth signaling pathways. Mutations in the members of the mitogen activated protein kinase (MAPK) signaling cascade have been observed in a wide variety of cancer types. Oncogenic mutations in the RAS family of genes (*HRAS*, *KRAS*, *NRAS*) are observed in 16% of all human cancers, with highly variable frequencies for specific cancer types (for instance *KRAS* mutations are observed in 33% and 61% of large intestine and pancreas cancers, respectively, while mutations in *NRAS* and *HRAS* have prevalence of at most 3% in these cancers) [28]. Mutations in the RAF family members are mostly observed in *BRAF* with 20% of cancer samples suffering an alteration, while mutations in *RAF1* (*CRAF*) are less common at about 2% across cancers. Melanoma genomes harbor *BRAF* mutations in 45% of the cases mostly being in a specific residue V600E [29].

Another deregulated signaling pathway is the PI3K/PTEN signaling pathway, and activation of *PI3K* via mutations and inactivation of *PTEN* via mutations or methylation are observed frequently in many tumors [30]. This pathway is especially

the most frequently mutated pathway in breast cancer, with mutations observed in more than 70% of the cases [31].

Insensitivity to antigrowth signals

Similar to growth factors, extracellular antigrowth factors ensure homeostasis in normal cells. Two mechanisms act in a complementary role one governed by the RB pathway and the other by the TP53 pathway. RB evaluates the external signals and decides whether or not to proceed with the proliferation cycle, whereas TP53 evaluates internal signals and halts the cell cycle or initiates apoptosis or senescence [2]. Thus, a cancer cell must inactivate these two tumor suppressor mechanisms in order to proliferate. As expected, alterations are observed in the members of both these pathways with large-scale NGS studies. For instance, RB pathway members are altered in 77% of primary glioblastomas tumors, and deletions in the *CDKN2A* and CDKN2B genes are the most frequently observed aberrations with 55% and in 53%, respectively [32]. Another mechanism deregulating RB signaling occurs through disrupting TGF-β signaling. In normal cells TGF-β prevents phosphorylation of RB thus suppresses cell proliferation, whereas cancer cells may inactivate TGF-β signaling through several genomic alterations. The TGF-β signaling pathway is known to be mutated in colorectal and pancreatic cancers [33], with mutation frequencies of 87%-27% (correlating with the hypermutation status) in colon cancers [20] and 100% in pancreatic cancers [34].

TP53 has a lot of critical roles in maintaining a healthy genome, and its inactivation is essential for most cancers not only for evading antigrowth signals but also for evading apoptosis. The genomic alterations observed in TP53 will be discussed in the next section.

Evasion of apoptosis

Cellular stresses including genomic instability and cellular hypoxia triggers programmed cell death by apoptosis in normal cells. Apoptosis is a defense mechanism for the organism to eliminate unhealthy cells from the system. However, cancer cells, despite being under cellular stress, evade from apoptosis and continue proliferation. Apoptotic pathways mainly have two components: "sensors" of cellular stress that receive signals either from extracellular or intracellular environment, and "effectors" that execute apoptosis upon receiving the signal from censors [2]. Cancer cells operate by deactivating these pathways either via over-expressing anti-apoptotic regulators, or by inactivating pro-apoptotic factors. The most common mechanism of apoptosis evasion is the loss of *TP53*. Somatic *TP53* mutations are observed in colorectal, head and neck, ovarian and pancreatic cancers at rates from 35% to 43%, while in leukemias the prevalence of *TP53* mutations is around 10%. The majority of the somatic *TP53* mutations are missense (73%), nonsense (8%) and frameshift mutations (9%), mostly clustered in the DNA binding domains (Data from IARC TP53 Database [35], R16 November 2012). It is also reported that in tumors with low mutation rates, alternative mechanisms take place for inactivating *TP53*, such as degradation of *TP53* by viral oncoprotein E in cervical

cancers [36], by nuclear exclusion of p53 in inflammatory breast cancer and neuroblastomas; and by amplification of *YEATS4* and *MDM2* genes in sarcomas [37]. Overall, NGS studies reinforced the role of *TP53* as the "guardian of the genome" [38], as it is one of the few genes that is altered at high frequencies across different cancers as its inactivation is required for the cancer cell to prosper [8]. Besides *TP53*, genes involved directly in apoptosis such as *CDKN2A*, *BCL2* and *MYC* are often mutated in cancers [5].

Limitless replication potential

Intrinsic mechanisms in normal cells limit the number of times a cell can replicate to 60-70 divisions, termed the "Hayflick limit" [39]. After this limit is reached, the cell enters irreversibly to a non-proliferative but viable state called "senescence". Knockdown of p53 and Rb in human fibroblast cells enable cells to replicate until they reach a state called "crisis" which is characterized by end to end fusions of chromosomes and massive cell death and a very rare occurrence (10^{-7}) of an immortalized cell without a replication limit [1]. The limiting factor of replication is found to be shortening of the telomeres, tandem hexanucleotide repeats located at the ends of chromosomes, protecting chromosome ends. These sequence repeats are shortened in each cell division, and the normal replication stops when the telomere sequences are critically short. In the majority of human cancers (85%) this limitation is overcome by expressing telomerase, the enzyme that extends the telomeres [40]. Alternatively, homologous recombination-mediated replication of telomeric DNA is also observed in a minority of cancers and is termed Alternative Lengthening of Telomeres (ALT) [41,42]. Recently, whole genome sequencing approaches have been used for assessing telomeric DNA content across 235 cases of pediatric cancers and marked differences in the telomere content were reported between tumor types [43]. Moreover, telomere gain had a significant association with high frequency of genomic rearrangements and somatic non-silent mutations, corroborating earlier remarks of telomerase being responsible for genomic integrity.

Sustained angiogenesis

Normal cells and cancer cells need to be close to blood vessels for obtaining nutrients and oxygen, and exerting metabolic waste and carbon dioxide. Sprouting new blood vessels, termed "angiogenesis," is not an intrinsic mechanism for normal cells and it is turned on only transiently in specific physiological processes such as wound healing and female reproductive cycling [2]. Thus angiogenesis is a hallmark that a cancer must achieve in order to grow. Angiogenesis is predominantly governed by the protein coded by the gene *VEGF*. Upregulation of *VEGF* by mutations in ras oncogenes [44] and VHL suppression are known mechanisms [45], and with pan-cancer NGS projects we also observe frequent mutations in VEGFR and FGFR family members. For instance in lung cancer VEGFR alterations (mutations and copy number aberrations) are found in 9% of the cases while FGFR mutations are observed in 19% [46].

Tissue invasion and metastasis

Cancer cells from the primary tumor site can invade and form tumors in other tissues of the organism in a process called "metastasis". This mechanism allows cancer cells to proliferate in a new environment that have more resources and nutrients than the primary site. The metastasis can be broken down into two phases: detachment from the primary tumor site and attachment to a new host site. Detachment in carcinomas involves the epithelial mesenchymal transition (EMT) program, which is observed in numerous processes in normal tissue development [2]. This process starts with the loss of E-cadherin, a cell-to-cell adhesion molecule, and a detached cancer cell, now with mesenchymal properties, may enter the bloodstream. EMT is a well-studied developmental program with known involvement of several transcription factors. Causal importance of these transcription factors are shown in experimental models, although the initiating event for EMT in cancer is largely unknown [2].

Attachment to a new host, or colonization, poses even a bigger enigma than detachment, as we still do not know what the genetic makeup allowing colonization is; or whether the cells require new genetic alterations to enable it. Multiple sequencing studies were performed to analyze the differences and similarities in the genetic makeup of primary and metastatic tumor cells in various cancers including pancreatic [47,48], breast [49], lung [50], renal [51] and colon [52]. For instance in pancreatic cancer, 64% of identified somatic mutations were identified as "founder" mutations; present in all metastasis sites as well as the primary site while 36% were "progressor" mutations, present in one or more metastases but absent in the primary site [47]. Several genes were identified with recurrent progressor mutations, however, the majority of them were altered in the primary site as well. Although there are indications for metastasis associated aberrations as exemplified by re-arrangements and mutations exclusive to the metastatic site in breast [49] and pancreatic cancers [53], no consistent genetic alterations were found to cause metastasis. Vogelstein *et al* proposes the possibility of explaining the lack of causal genetic alterations for metastasis simply by stochasticity of the cancer genome evolution: the bigger and more advanced the tumor, the more probable that a cancer cell would find a hospitable environment fitting to its genotype [5]. Thus the clonal diversity of the tumor enabled by the genomic instability, may allow the invading cells to thrive in a new environment.

Reprogramming energy metabolism

Cancer cells can change their energy metabolisms from mitochondrial oxidative phosphorylation to glycolisation even in the presence of oxygen, thus leading to a state called "aeorobic glycolisis". This process is 18-fold less efficient than oxidative phosphorylation however it has the added benefit of biosynthesis of macromolecules that is used for formation of new cells; necessary for the actively proliferating cancer cells [2].

Many of the genes that are altered for enabling the above-mentioned hallmarks also take roles in regulating energy metabolism of the cell such as *MYC*, *NFKB*, *AKT*, *EGFR* and *TP53* [2,54]. However, mutations targeting the genes encoding mitochondrial enzymes have also been found. Rare mutations in the genes *SDHB* and *FH*, which encode tricarboxylic acid (TCA) cycle enzymes, have been found in familial paraganalioma and papillary renal cell cancer [54]. The TCA cycle is used less by cancer cells, and the aforementioned mutations force a switch to aerobic glycolisis [54]. More frequently, mutations in *IDH1* and *IDH2* genes have been reported first in gliomas [55] and then in acute myeloid leukemias[56] and chondrosarcomas [57]. These mutations have been selected for altering the energy metabolism, however their roles in tumorigenesis remain unclear [58].

Evading immune destruction

The immune system defends the organism against a plethora of threats including bacteria, viruses, and as recent research indicates, against tumor formation and progression [2]. Thus overriding the immune system is actually a hallmark achieved by every full-blown tumor. In prevention of the cancer, immune system has mainly three operational duties: suppression of viral infections (to reduce incidence of cancers with viral origin), elimination of pathogens (prevent or shorten the inflammation, which benefits tumorigenesis) and immunosurvellience (identification and destruction of transformed cells before they establish malignancy) [59]. Tumor cells evade the immune system either by avoiding recognition or by developing mechanisms to escape immune-mediated killing [59]. Studies in cell culture and mouse models have showed that loss of interferon mediated signaling due to disruptions in the JAK-STAT signaling cascade or loss of interferon regulated genes such as MHC molecules, *TAP1*, *PSMB9* have been shown to contribute to evasion from immune recognition [60]. To this end, mutations in the *HLA-A* gene in lung cancers have been observed with a possible contribution to achievement of this hallmark [61]. *HLA-A* encodes the alpha chain of the MHC molecule, which is expressed at the cell surface and mediates interaction with immune cells. Loss of MHC surface expression might lead to evasion of the tumor cell from the immune system and inactivating mutations in *HLA-A*, observed in 3% of lung cancers, might underlie this phenomenon.

Gene regulation in cancer

Another view on the cancer processes can be provided through molecular functions governing them. All the above-mentioned processes are mediated through cascades of events in the intricately woven molecular networks in the cell and transcriptional regulation is at the heart of it. Deregulated transcriptional control results in changes in the gene expression levels that in turn affect various cancer processes such as apoptosis and cell cycle. Indeed many of the classical cancer genes are transcription factor or signaling molecule-encoding genes, whose misregulation propagate through their downstream effectors and disrupt the biological processes they perform under

normal conditions. For instance, the *TAL1* transcription factor is over-expressed in almost half of the T cell acute lymphoblastic leukemia (T-ALL) cases and it drives aberrant proliferation, differentiation and survival [62]. Mutations in signaling molecules affect the downstream activity of transcription factors. For instance alterations in the MAPK signaling cascade (as mentioned in "self sufficiency in growth signals") results in activation of the transcription factors such as *MYC*, *CREB* and *FOS*, which in turn alter the cell cycle process.

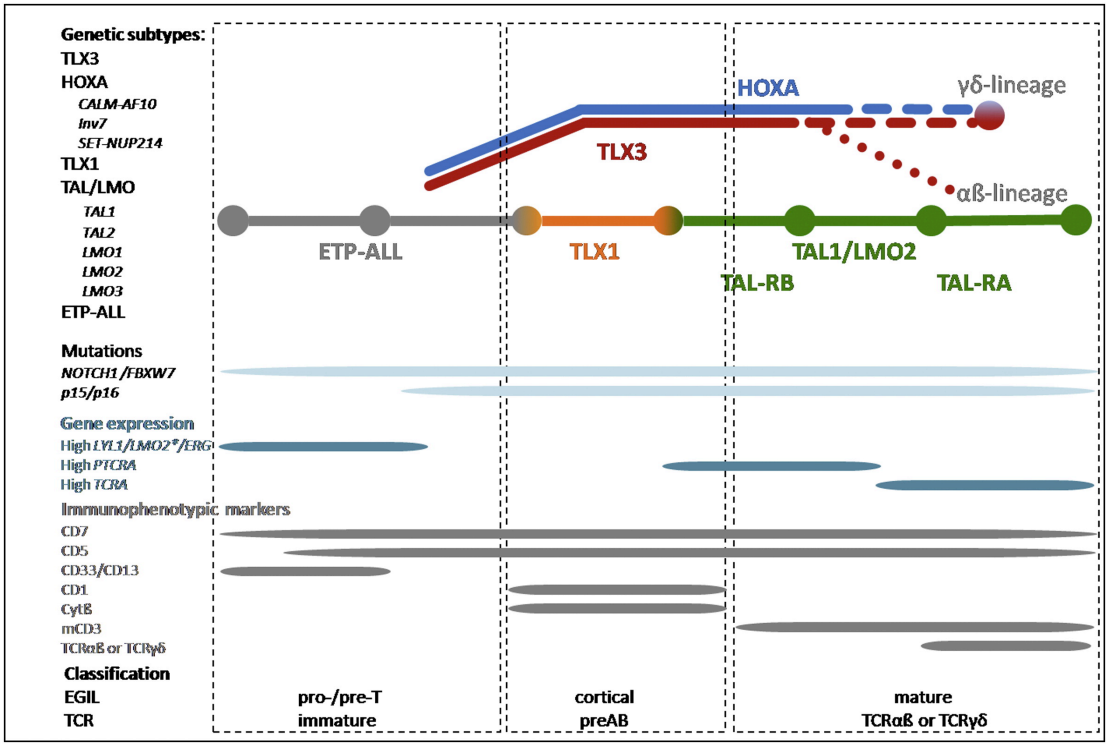
In addition to genes with specific roles in cancer processes, NGS studies identified driver genes that have broad functions such as chromatin modification, DNA methylation and RNA splicing. Indeed, 25% of the true cancer drivers identified by Vogelstein *et al* are chromatin modifiers [5]. Initially it was unclear if these changes had a causal relationship with cancer, however their recurrence across a wide range of tumors ensured their *driver* status [63]. These genes are involved in transcriptional regulation in an indirect manner by changing the accessibility of the genome (for transcription) through chromatin modifications and DNA methylation, or by disrupting the splicing machinery, which regulates the amount and type of the mature mRNAs. For instance, mutations in several nucleosome remodeling genes such as *ARID1A*, *SMARCA1* and *SMARCA4* are found in several cancers including colorectal [20], ovarian [64], and renal cell [65] carcinomas as well as lymphomas [66]. These mutations result in inactivation of the gene, implying the genes that are actively transcribed in these cancers as candidate genes that are affected by the misregulated nucleosome remodeling program [67]. Indeed, ARID1A containing complexes have been shown to repress c-Myc during differentiation [68], pointing to another way of activating c-Myc in tumorigenesis with loss-of-function mutations in *ARID1A* . Similarly, mutations in splicing factors have been identified in chromic myeloid leukemias with *SF3B1* mutations in 10-15% of the cases, then in solid tumors with mutations in *U2AF1*, *SF3B1*, *U2AF2* and *PRPF40B* [63], however the specific targets of spliceosome defects in tumorigenesis remain to be elucidated.

MOLECULAR PATHOGENESIS OF T-CELL ACUTE LYMPHOBLASTIC LEUKEMIA

In this thesis, T-cell acute lymphoblastic leukemia (T-ALL) genomes and transcriptomes will be explored using next generation sequencing technologies in order to detect driver mutations and gain insight into pathogenesis of these malignancies. This chapter provides an overview on the current status of known genomic aberrations and molecular mechanisms in operation in these tumors.

T-ALL is a hematological cancer arising from malignant transformation of lymphoid progenitors. Chromosomal translocations and cooperating point mutations equip the progenitor T-cell with diverse hallmark capabilities. At the core of genetic alterations underlying T-ALL are aberrations resulting in a differentiation block and disrupted NOTCH signaling (**Figure 1**). Additional mutations contribute to tumorigenesis by altering cell cycle, growth and proliferation mechanisms (**Table 1**).

Figure 1. Schematic overview of T-ALL genetic subgroups [69]



Differentiation block

A block in differentiation is a hallmark of T-ALL and is usually achieved by chromosomal translocations resulting in the activation of differentiation specific transcription factor (TF) genes. These translocations typically involve juxtaposition of regulatory elements of T-cell receptor (TCR) genes that are located on chromosome 7q34 (*TCRB* and *TCRG*) and chromosome 14q11 (*TCRA* and *TCRD*)

to TFs most of which are not expressed in normal T-cell development with the exceptions being *TAL1* and *LYL1*. In addition, alternative genetic rearrangements can activate these TFs such as deletions in chromosome 1p32 and 11p13 leading to *TAL1* [70] and *LMO2* overexpression [71], respectively and gene duplication leading to *MYB* overexpression [72]. Aberrant expression of these TFs results in the accumulation of immature cells in defined differentiation blocks and this is also reflected in the transcriptional profile of the tumor cells. Indeed, distinct molecular subtypes caused by these alterations can be identified by gene expression signatures [73].

NOTCH signaling

Apart from translocations, constitutively active NOTCH signaling is at the core of T-ALL pathogenesis. During normal T-cell development, *NOTCH1* participates in cell fate specification and NOTCH signaling is required throughout different developmental stages. Mutations in *NOTCH1* are observed in more than 60% of T-ALL cases [74] throughout all distinct subtypes causing hyperactive NOTCH signaling. Interestingly, activation of the NOTCH signaling pathway is not only due to *NOTCH1* mutations but also due to loss function mutations in *FBXW7*, which takes part in NOTCH1 degradation [75]. Oncogenic NOTCH1 is involved in plethora of activities such as cell growth, proliferation and survival with interaction partners including *MYC* and *NFKB.*

Cell cycle defects

Deregulation of cell cycle in T-ALL occurs through inactivation of *CDKN2A* and *CDKN2B* genes [76]. These genes are located in chromosome 9p21, and this region is deleted in 70% of the cases [77]. *CDKN2A* encodes two tumor suppressor proteins p16 and p14, whereas *CDKN2B* encodes for p15. P16 and p15 block cyclin D–CDK4/6 complexes that positively regulate cell cycle through phosphorylation of RB1, and p14 inhibits MDM2 which induces the activation of TP53. Thus aberrations in CDKN2A and CDKN2B implicate not only deregulation of *RB1* but also *TP53*.

Cell growth and survival defects

Another class of altered TFs affect cell growth and survival pathways. Although a lot of recurrent aberrations are observed in this class of genes, their contribution to T-ALL tumorigenesis is still an active research area [76]. MYC is among these genes and it is a prominent oncogene not only in T-ALL but across many other human cancers taking part in cell growth and proliferation [78]. In T-ALL, the *MYC* oncogene is activated directly by *NOTCH1*, however in 1% of the cases *MYC* activation can be realized by a t(8;14)(q24;q11) translocation [79]. The remaining alterations in this class are loss of function events affecting tumor suppressor genes either through inactivating mutations or deletions. The majority of these genes are important regulators in hematopoiesis and alterations are observed in other hematological malignancies exemplified by *ETV6* alterations found in B-ALLs, and *RUNX1* inactivating mutations in AML [76].

Alterations in signaling pathways

In addition to the above-mentioned direct regulators of cell cycle, growth, proliferation and differentiation, the signaling pathways regulating these processes are frequently mutated in T-ALL. Primarily AKT, RAS and JAK-STAT pathways are deregulated through genetic alterations. *PTEN* acts as a negative regulator of the AKT pathway, and inactivating mutations and deletions lead to uncontrolled proliferation in the tumor cells [80]. RAS signaling, on the other hand, is disrupted through mutations in *NRAS* and *NF1*, which encodes a negative regulator of RAS signaling pathway. And finally, the JAK-STAT pathway is deregulated due to mutations in *JAK1* and *JAK3*, or translocation involving *JAK2*. Additionally, a gain of function mutation in *IL7R* leads to constitutive JAK/STAT signaling [76].

Mutations in chromatin remodelers

Polycomb repressive complex 2 (PRC2) is responsible for repressive marking of the chromatin and is composed of four proteins encoded by *SUZ12*, *EZH2*, *EED* and *RBBP4*. Two of these genes, *SUZ12* and *EZH2*, have been reported to be mutated in T-ALL up to 25% of the cases and functional evidence points to a tumor suppressor role for these mutations [81,82]. Another chromatin modifier, *PHF6* is mutated in 16% of pediatric and 38% of adult T-ALL cases. Notably, mutations are observed almost exclusively in male patients, resulting in a hemizygous mutation/deletion [83].

Table 1. Recurrent genetic alterations in T-ALL. Adapted from [76]

Category	Gene target	Genetic rearrangement	Frequency
Differentiation impairment (through TF aberrations)	*TAL1*	t(1;14)(p32;q11) t(1;7)(p32;q34) 1p32 deletion	3% 3% 16-30%
	TAL2	t(7;9)(q34;q32)	1%
	LYL1	t(7;19)(q34;p13)	1%
	BHLHB1	t(14;21)(q11.2;q22)	1%
	LMO1	t(11;14)(p15;q11) t(7;11)(q34;p15)	1% 1%
	LMO2	t(11;14)(p13;q11) t(7;11)(q34;p13) 11p13 deletion	6% 6% 3%
	LMO3	t(7;12)(q34;p12)	<1%
	TLX1	t(11;14)(p15;q11)	5%-10% & 30%
	TLX3	t(11;14)(p15;q11)	20% & 5%
	HOXA	Inv(7)(p15q34) t(7;7)(p15;q34)	3% 3%
	HOXA (CALM-AF10)	t(10;11)(p13;q14)	5%-10%
	HOXA (MLL-ENL)	t(11;19)(q23;p13)	1%
	HOXA (SET-NUP214)	9q34 deletion	3%
	NKX2.1	inv(14)(q13q32.33) t(7;14)(q34;q13)	5%

	NKX2.2	t(14;20)(q11;p11)	1%
	MYB	t(6;7)(q23;q34) Gene duplication	3% 8%
NOTCH1 pathway	*NOTCH1*	t(7;9)(q34;p13)	<1 %
		Activating mutation	>60%
Cell cycle defects	*FBXW7*	Inactivating mutation	8%-30%
	CDKN2A/2B	9p21 deletion methylation	70%
	CCND2	t(7;12)(q34;p13) t(12;14)(p13;q11)	1%
	RB1	13q14 deletion	4%
	CDKN1B	12p13 deletion	2%
Cell growth and survival defects (through TF aberrations)	*MYC*	t(8;14)(q24;q11)	1%
	WT1	Inactivating mutation/deletion	10%
	LEF1	Inactivating mutation/deletion	10%-15%
	ETV6	Inactivating mutation/deletion	13%
	BCL11B	Inactivating mutation/deletion	10%
	RUNX1	Inactivating mutation/deletion	10%-20%
	GATA3	Inactivating mutation/deletion	5%
Signal transduction	*PTEN*	Inactivating mutation 10q23 deletion	10% 10%
	NUP214-ABL1	Episomal 9q34 amplification	4%
	EML1-ABL1	t(9;14)(q34;q32)	<1%
	ETV6-ABL1	t(9;12)(q34;p13)	<1%
	BCR-ABL1	t(9;22)(q34;q11)	<1%
	NRAS	Activating mutation	5%-10%
	NF1	Inactivating mutation/deletion	3%
	JAK1	Activating mutation	4%-18%
	ETV6-JAK2	t(9;12)(p24;p13)	<1%
	JAK3	Activating mutation	5%
	FLT3	Activating mutation	2%-4%
	IL7R	Activating mutation	10%
Chromatin remodeling	*EZH2*	Inactivating mutation/deletion	10%-15%
	SUZ12	Inactivating mutation/deletion	10%
	EED	Inactivating mutation/deletion	10%
	PHF6	Inactivating mutation/deletion	20%-40%

UNDERSTANDING CANCER GENOMES THROUGH NEXT GENERATION SEQUENCING

Cancer is a disease of mutations. The first indication that cancer was a genomic disease came from Boveri [84] and von Hansemann [85] in the beginning of the twentieth century. They observed unusual chromosomal aberrations in cancer cells and postulated that cancers are abnormal clones of normal cells caused by abnormalities in the hereditary material. In 1973, the Philadelphia chromosome (the translocation between chromosomes 9 and 22) was identified as a first example of a genomic abnormality associated with a particular cancer type [86]. A decade later, the first somatic mutation in a human cancer was found: a point mutation in the *HRAS* gene[87,88]. The completion of the human genome sequencing project further fueled the cancer genomics field both by providing the technology for sequencing DNA in an automated fashion and by supplying the backbone in the form of the annotated human genome reference sequence. Despite the high cost and low throughput of sequencing technology, a number of studies took the task of identifying disease causing mutations in cancer either by sequencing a limited set of selected exons in a large cohort [6,7] or all known coding exons on a limited number of samples [8,9]. However, it were the next generation sequencing (NGS) technologies that shifted the focus from targeted small-scale approaches to comprehensive genome-wide approaches. The availability of NGS platforms for research groups worldwide has resulted in a remarkable advancement in our understanding of the mutational landscapes of many human cancers. Studies examining genomic aberrations in targeted genomic regions and whole genomes have been reported for a large number of tumor samples (**Table 2**). Moreover, systematic cancer genome sequencing projects have emerged: first the Cancer Genome Project (CGP) [92] in the United Kingdom and then The Cancer Genome Atlas (TCGA) [93] in the USA; both aiming to sequence several major cancer types. The International Cancer Genome Consortium (ICGC) was then founded to coordinate data generation and analysis, and currently 52 cancer types are being sequenced and analyzed under the ICGC umbrella [94]. Comprehensive genomic analysis have been released for various cancer types by these consortia as well as individual research groups, and Table 2 lists a few examples. Today, more than 1% of all human genes are implicated in cancer via mutations and overall, more than 290 000 somatic mutations in 41 main cancer types have been identified (data extracted from COSMIC v66) [95].

Converting genomic data into biological knowledge

The massive amounts of data produced by next generation sequencing (NGS) platforms enable comprehensive characterization of mutational landscapes of cancer genomes, however this becomes possible only after addressing the computational and statistical challenges in the data analysis. This section will lay out the technical

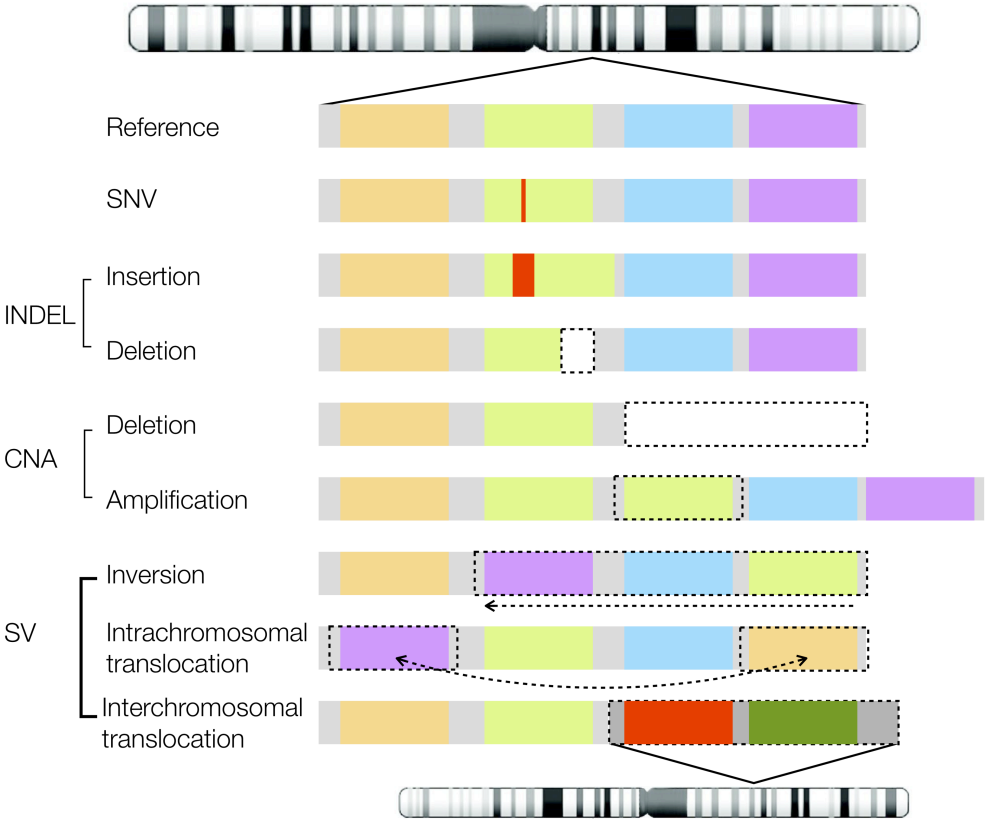
aspects of cancer genome sequencing and provides the state of the art for the analysis approaches.

Table 2. Overview of cancer genomics studies that appeared during this PhD

Tumor type	Cases	Sequencing type	Reference
Acute Myeloid Leukemia	200	Whole genome & exome	TCGA [96]
Breast cancer	507	Exome	TCGA[97]
Breast cancer	100	Exome	[98]
Breast Cancers	65	Whole genome & exome	[99]
Burkitt Lymphoma	60	Whole genome & exome	[66]
Chronic Lymphocytic Leukemia	105	Exome	[100]
Colorectal Cancer	224	Exome	TCGA [20]
Colorectal Cancer	74	Whole genome & exome	[101]
Diffuse large B-cell lymphoma	94	Exome	[102]
Diffuse large B-cell lymphoma	55	Exome	[103]
Endometrial cancer	248	Exome	TCGA [104]
Head and Neck squamous cell carcinoma	74	Exome	[105]
Kidney renal clear cell carcinoma	417	Exome sequencing	TCGA [65]
Lung Adenocarcinoma	183	Whole genome & exome	[106]
Lung squamous cell carcinoma	178	Exome	TCGA [61]
Melanoma	147	Exome	[107]
Melanoma	121	Exome	[108]
Ovarian Carcinoma	316	Exome	TCGA [109]
Pancreatic Adenocarcinoma	99	Exome	[110]
Pediatric Medulloblastoma	60	Whole genome & exome	[111]
Pediatric Neuroblastoma	87	Whole genome	[112]
Prostate Cancer	112	Exome	[113]
T-cell acute lymphoblastic leukemia	67	Exome	[114]

The genomic alterations that can be identified by NGS technologies consist of differences in the sequence or in the arrangement of sequence blocks compared to the reference genome. These include single nucleotide variations (SNV), small insertions and deletions (INDEL), copy number aberration (CNA) and structural variations (SV) (**Figure 2**). SNVs involve changes in single bases, while INDELs span multiple base pairs up to 1 kilobase [115]. When they occur in the protein coding regions, they might change the encoded amino sequence (missense mutation), introduce stop codons (nonsense mutations) or, in the case of INDELs, change the reading frame (frameshift mutations). These alterations could influence the activity of the encoded protein or result in loss of the protein product altogether. On the other hand, when they occur in the non-coding genome they can disrupt the splicing patterns of the gene via splice site mutations (causing intron retention or exon skipping events) [116] or alter gene transcription via mutations in the regulatory regions [117,118].

Figure 2. Schematic representation of genomic variations detected by NGS

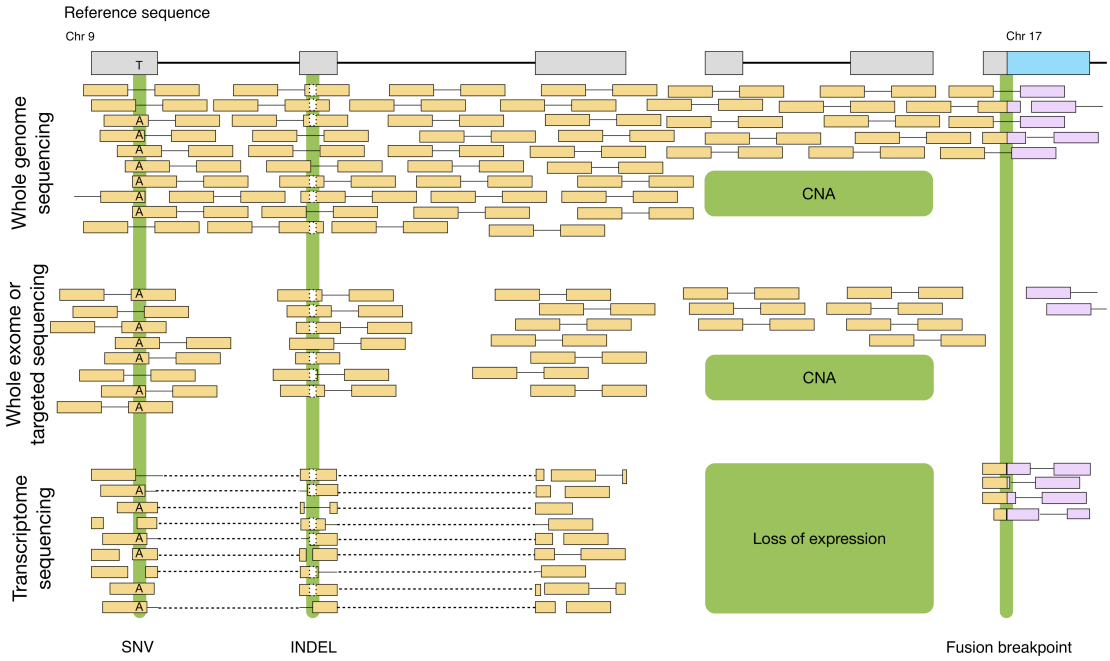


CNAs include amplifications and deletions of large chromosomal regions as well as focal gains and losses. These alterations could result in gene duplications or deletions, thus influencing gene dosage or may lead to the complete loss of genes located within them. SVs disrupt the composition of the genome via rearrangement of genomic material and could be observed in the forms of inversions or translocations. Inversions are rearrangements in which DNA segments are re-inserted in to the genome in a reverse orientation. Translocations, on the other hand, involve the interchange of genetic material between two distinct genomic regions either within the same chromosome (intrachromosomal) or between different chromosomes (interchromosomal). Both inversions and translocations could result in effects such as disrupting the regulatory sequences that control gene expression or creating genetic rearrangements like gene fusions.

The range of genomic aberrations that can be identified in an NGS study depends on the sequencing approach used (**Figure 3**). Whole genome sequencing (WGS) provides the most comprehensive view on the genome allowing identification of all mutations, rearrangements and copy number changes effectively and simultaneously. Variations of this approach can allow sequencing defined regions in a genome, termed 'targeted sequencing'. This targeted approach involves enrichment of a library for the desired target regions using PCR or hybridization based approaches, and subsequent sequencing of this library. Targeted sequencing strategies could be used for sequencing a set of interesting genes or the entire coding sequences of the

genome (exome sequencing) for a complete characterization of the mutational profile of the samples, as demonstrated in Chapter 3 of this thesis. In addition, these approaches can be valuable to confirm mutations found by other NGS approaches or for uncovering the clonal architecture of a tumor by deep sequencing the potential driver genes or mutations [119].

Figure 3. Different types genomic aberrations can be identified with different NGS approaches. Adapted from [120]



NGS approaches can also be used for characterizing cancer transcriptomes. Transcriptome sequencing, which is also called RNA-seq, focuses on sequencing the cDNA, obtained from the mRNA, total RNA or other RNAs such as micro-RNAs. RNA-seq provides an efficient way of measuring gene expression across whole transcriptome since the number of RNA-seq reads that map to a particular region can be used as an abundance measure. This feature also allows measuring expression at the level of transcripts and permits discovery of alternative transcript events (ATEs). In addition, RNA-seq can be used to inquire mutations in the expressed genes and identify fusions.

Another application of NGS includes epigenome sequencing, which is not discussed in this thesis. Briefly, these approaches are used for characterizing the genome-wide DNA methylation and histone modification profiles as well as assessing chromatin accessibility and include techniques such as ChIP-seq, DNase-seq, and MeDIP-seq [121].

Bioinformatics pipelines for NGS analysis in cancer genomics

NGS data analysis is executed through 'pipelines' that include data processing steps performed with algorithms and methods specific for the biological question and the sequencing experiment at hand. A typical NGS analysis pipeline includes preprocessing of the sequence reads to remove adapter sequences and low quality

reads, alignment of the reads to a reference sequence, post-alignment modifications, detection of the genomic variants and identification of the driver aberrations (**Figure 4**).

Alignment of the reads to the reference sequence

After adapter trimming and ensuring the overall base quality of the sequence reads is acceptable (Table **3.A-B**), the reads are aligned to the reference genome sequence. Tens to hundreds of millions of reads need to be mapped to the reference sequence, typically the genome or transcriptome. Due to the read lengths and the repetitive nature of the human genome, some reads might map to several places in the reference sequence. Furthermore, the presence of sequence mistakes along with genuine mutations makes it harder to find the most likely source of origin in the reference sequence. Several mapping algorithms have been developed specifically addressing these challenges (**Table 3.C**). The general strategy of these algorithms is to follow a two-step mapping procedure starting with the detection of the most probable location of the read in the reference sequence with fast heuristic approaches (such as hash-table based or Burrows Wheeler transform methods), followed by slow and accurate mapping of the read to the reference sequence (such as Smith-Waterman) [122].

Figure 4. A high level overview of NGS analysis

To further increase the accuracy of the alignment, a number of post-alignment modifications might be performed consisting of three steps: removal of duplicate reads from the alignment, local re-alignment around INDEL sites and base quality score re-calibration [123] (**Table 3.D**). The first post-alignment modification is the removal of duplicate reads that originate from the PCR amplification step in NGS library preparation protocols. The PCR amplification step might be biased towards short fragments or fragments with lower GC composition, and this might lead to artificial differences between coverage levels. Moreover, if this biased amplification occurs for a genomic material with an early PCR mistake, it might cause a false positive variant prediction in the downstream analysis. To prevent this systematic bias, pre or post alignment duplicate removal is often implemented. A pre-alignment duplicate removal strategy involves removing sequence reads that are identical in the raw *fastq* file. A post-alignment procedure incorporates the location of the read in the genome and removes the reads that have identical beginning and end coordinates.

The second step in this framework is the local re-alignment of the sequences around INDEL sites. Presence of small insertions and deletions in the sequence might pose

problems in the mapping step as the aligners might introduce mismatches instead of gaps or insertions. Since the aligners handle each sequence read individually, a consensus variant sequence is hard to generate at true INDEL sites. These sites are often mapped with multiple mismatches instead of the INDEL, or even if the INDEL is mapped correctly the read harbors mismatches in the vicinity of the INDEL [123]. To tackle this problem, GATK Indel Realigner evaluates the sequences around known or predicted INDELs, or sites with a cluster of mismatching bases with a multiple sequence alignment approach. This step cleans the regions with such mapping artifacts, improves INDEL calling in the downstream analysis and decreases the false positive prediction rate of SNVs.
The last step of the post-alignment modifications is the recalibration of the base quality scores. The sequencer assigns a quality score indicating the accuracy of a base call. However, it has been shown that the quality scores are actually not accurate and affected by a number of factors such as sequencer chemistry, machine cycle (the position within the read), and di-nucleotide context (preceding and the current base) [123-125]. The recalibration ensures that the base quality score reflects the probability of a sequencing error and not co-vary with other factors. The GATK implementation calculates correction factors for each class of covariates, however other implementations exist based on logistic regression [124] or Hidden Markov Models [126].

Genome Variation Discovery

SNV discovery

The post-alignment modifications, described above, provide well-mapped, realigned and re-calibrated reads ready for the identification of variants. Variant calling can then be performed using a variety of tools as listed in Table 3.E. The challenge is to distinguish true variants from alignment and/or sequencing mistakes. In its most simple form, variant calling can be done by counting alleles at each site and using simple cut-offs for deciding if a base is variant or reference. Such an approach was followed by VarScan in which genotype calls are done in positions above certain coverage and quality, then a SNV prediction is made if there is a variant base above a certain variant allele frequency (additionally, a p-value is calculated based on Fisher's exact test for the read counts supporting the reference and variant alleles versus the expected distribution of alleles based on sequencing error) [127]. These simple methods work well with high coverage sequencing depths, however with low sequencing depths (e.g., <20X) the fixed cut-offs on variant allele frequency thresholds leads to under-calling of heterozygous variants [128]. Furthermore, the use of fixed cut-offs on the sequencing metrics (depth of coverage, variant allele frequency, quality scores) leads to information loss. Probabilistic methods, on the other hand, incorporate the quality scores along with other metrics from the sequencing, and contain uncertainty in the prediction. Most algorithms converge on a Bayesian approach for this task. Briefly, these methods calculate the posterior

probability of each genotype given the read data for a particular site, using the Bayesian formulation:

$$p(G \mid D) = \frac{p(G)p(D \mid G)}{p(D)}$$

where D represent the data (the pileup of reads and qualities in a given base) and G represent the genotype. The genotype likelihood can be calculated as follows for a diploid genome, assuming the reads are independent:

$$p(D \mid G) = \prod p(b \mid G) \text{ where } p(b \mid G) = p(b \mid \{A_1, A_2\}) = \frac{1}{2} p(b \mid A_1) + \frac{1}{2} p(b \mid A_2)$$

with A_1 and A_2 being the two alleles of a diploid genome in a given position. Generally the genotype with the highest probability is called and the probability value is reported as a measure of confidence [128,129].

INDEL discovery

Detection of INDELs is complicated due to mapping problems of the sequence reads containing INDELs (as mentioned in the beginning of this section). An optimal combination of a mapping and an INDEL detection method is essential for accurate and specific INDEL predictions. Dindel [130], a Bayesian approach for calling INDELs, is built on this premise : candidate INDELs are collected from the alignment file and candidate haplotypes (representing the reference sequences with the imputed candidate INDEL) are used for remapping the reads in the regions with candidate INDELs. It had the best performance when compared to three other INDEL callers (GATK, SAMTools mpileup, VarScan) [131] , however it is suitable only for Illumina analysis. Other methods range from simple heuristic filters (eg. VarScan [127,132]) to pattern growth approaches (eg. Pindel [133]) (**Table 3.F**), although often parameter optimization is required to obtain reliable INDEL predictions [131].

Copy Number Aberration Discovery

Although CNA detection is more common practice with arrays or WGS, methods are being developed to detect CNAs using exome sequencing as well. The discrete nature of the reads in Exome-seq is a challenge for CNA detection. Moreover, read depth is not exclusively correlated with copy number status since biases in sequence capture as well GC content affects the sequencing depth. A variety of methods are available now (**Table 3.G**), mostly employing Hidden Markov Models for identification of CNAs.

Transcriptome Variation Discovery

Gene expression changes do not actually constitute a class of genomic variation, however a change in gene expression may coincide with an underlying genomic aberration, thus this data can be used to enhance understanding the observed alterations. Gene expression levels can be estimated from the number of reads

mapping to each gene in the RNA-seq data. The first step is to count the number of aggregate reads in a given gene. Intuitively, counting the number of reads overlapping a defined gene region seems trivial, however there a number of subtleties to be addressed such as tackling the reads that are not uniquely assigned to a gene. For this reason a number of tools have been developed allowing user flexibility to handle these problems (**Table 3.H**). Next, the expression values need to be normalized to remove systematic biases. The sources of bias include between sample differences such as library size (sequencing depth) [134], and within sample effects due to differences in gene length [135] and GC content [136]. Several normalization metrics and methods were developed over the last few years addressing the removal of these biases such as Upper Quantile (UQ) [137], DESeq normalization [138], Trimmed Mean of M values (TMM) [139], and Reads Per Kilobase per Million mapped reads (RPKM) normalization [134]. Additionally, normalization methods based on the use of housekeeping genes [137] and GC-content bias [140] have also been proposed. Dillies *et al* conducted a comprehensive evaluation on different RNA-seq normalization techniques and concluded that DESeq and TMM methods provide reasonable normalization based on the qualitative characteristics and impact of the differential expression[141]. Both of these methods are based on the hypothesis that most of the genes are not differentially expressed, and they both calculate a correction factor for removing the bias of library size. Following the normalization, exploratory analysis can be conducted using clustering or classification algorithms in akin to microarray analysis. Additionally, differential expression tests can be performed using parametric or non-parametric approaches [142]. Parametric approaches assume that each expression value is drawn from a particular distribution such as Poisson [143], Negative Binomial [138,139,144-147], or Beta Binomial [148] while non-parametric models estimate the noise from the count data.

RNA-seq data can provide a comprehensive view on a transcriptome at hand by allowing identification and quantification of all expressed exons and transcripts. However, several factors make this analysis challenging: (1) non-uniform coverage across the transcriptome with some transcripts supported by a few reads and some with extensive coverage and (2) shared exons between different transcripts [149,150]. Several methods have been developed addressing these challenges (**Table 3.I**). Two of these methods, Scripture [151] and Cufflinks [152,153] perform *de novo* transcriptome assembly using spliced reads. Both methods follow a similar approach of creating an assembly graph and parsing the graph for paths to infer possible transcripts. The main difference between the two methods lie in the way of parsing the graph: Scripture focuses on maximum sensitivity and reports all the transcripts compatible with the read structure while Cufflinks aims at maximum precision and reports the minimal number of compatible isoforms [150]. These methods lay out the transcriptome from a given RNA-seq experiment, leading to identification of both known and novel isoforms. Estimation of expression levels of these entities is tackled by other methods developed to overcome the 'read assignment uncertainty' problem. Alexa-seq focuses on the reads that exclusively map to a single isoform and ignore

the ambiguous reads [154]. However, other approaches like Cufflinks and MISO attempt to solve this problem by constructing likelihood functions that estimate the expression with the given read structure [152,153,155].

Finally, RNA-seq data can be used to obtain gene fusions (**Table 3.J**). Methods that rely on the paired-end read structure of the RNA-seq data identify fusion genes using discordant reads, which are read pairs that have a significantly different inter distance compared to the rest of the read pairs; and split reads, which span the fusion boundary. deFuse follows such a strategy: the algorithm first finds the discordant reads and generates putative fusions events [156]. The algorithm subsequently searches for split reads spanning the putative fusion boundary and implements a set of filters to generate final predictions. Another class of methods fragment the reads and map these to the reference sequence. Then the mapped fragments are used for generating putative fusion events. Algorithms employing this strategy are MapSplice [157], FusionMap [158] and FusionFinder [159]. A third class of methods involve a combination of these two methods: discordant reads are used for creating putative fusion references, then the reads that do not map initially are fragmented and mapped to this new reference sequence. Tophat-fusion[160] and ChimeraScan [161] follow this strategy.

Identification of somatic mutations and driver genes

It is the *somatic* mutations that eventually leads to cancer [1] either in a stepwise manner leading to gradual accumulation or through a crisis-driven manner (for example, chromothripsis, chromoplexy, kataegis) with the introduction of a high number of mutations in a relatively shorter time scale [6]. After obtaining high quality genomic variants from NGS data, the next step is to identify their somatic status. When a matching normal sequence is available for a sequenced tumor sample, somatic mutation detection can be performed. Subtracting the variants that are called in the matched germline from the tumor sample variants can be seen as a simple and straightforward approach, yet it is biased towards regions that are not covered in the either of the samples, or for tumor variants with low allele frequencies [162]. Instead, somatic SNV detection algorithms have been developed for joint analysis of matched tumor and normal samples as listed in Table 3.K. These methods mostly rely on a Bayesian framework to model the genotypes, (as explained in the SNV discovery section) however simple models with heuristic filters do exist (VarScan2 [132]). A comparison between Bayesian framework models and Varscan2 revealed that high probability candidate somatic mutations from VarScan2 were also identified with other methods, while the low probability predictions suffered from a high rate of germline false positives, and the method was unable to detect somatic variants with low variant allele frequencies (<0.2) [163]. Assessment of different implementations of Bayesian framework models demonstrated that they perform equally well with high sensitivities (>99%) for coverage above 30X and variant allele frequency above 0.4, however as the variant allele fraction decreases the sensitivity

decreases as well, and at 0.1 allelic fraction MuTect was the most sensitive method with 53.2% sensitivity [162].
In the absence of the matched germline data, common population variants can be used to approximate somatic variants. HapMap [164] and 1000 genomes [165] projects have identified over 55 million population variants. These variants are detected in the healthy population with minor allele frequencies of 5% and 1%, respectively. However, the use of these databases requires caution as a small percentage of pathogenic variants are infused to these databases.

Pathogenesis of cancer is driven by driver genetic changes which confer selective growth advantage to the cancerous cell [5]. However, a cancer cell does not harbor only driver events but also 'passenger' events, which do not contribute to malignant transformation and occur due to stochastic mutation processes amplified by the genomic instability.
Methods dealing with driver mutation identification can be broadly divided into two categories: mutation significance based methods and functional consequence based methods (**Table 3.L**). Even though they evaluate the problem in different ways, the methods in these two categories are often used in combination. The mutation significance based methods focus on the SNV and INDEL predictions and work on the assumption that the driver genes harbor more (ie. frequency based methods) or have different mutation patterns compared to passenger mutations. Among these significance based methods GenomeMusic [166] and MutSigCV [167] assess the mutation significance based on the mutation frequency and calculate the background mutation rate from the matched germline samples or from the silent and non-coding mutations from the same sample, respectively. Sjoblom *et al* employ a different strategy for estimating background mutation rate and incorporate nucleotide type and context into the calculation: they divide mutations into categories according to the nucleotide context the mutation resides in (ie. A, C, T, G mononucleotide sites; CG, TC and GA dinucleotide sites) and estimate a distinct background mutation rate for each mutation type [90].
Functional consequence based methods employ the impact the genomic variant has on the function of the gene. A variant can have a gain-of-function, loss of function or neutral effect on the protein. Gain-of-function events result in an aberrant or ectopic activity of the gene it is observed, while loss-of-function events result in loss of activity the gene performs. Typically, these two events are of interest in cancer studies, and can be collectively termed as 'protein altering events'. Methods such as Variant Effect Predictor [168], SeattleSeq [169], and Annovar [170] assess the functional impact of the variant on the protein product and can be used for detecting protein-altering events, while SIFT [171] and PolyPhen [172] employ additional information such as evolutionary conservation, the location of the variant on the 3D structure of the protein and chemical similarity between the known and novel (mutation imputed) protein. Furthermore, there are also methods that incorporate cancer specific information on top of the sequence information such as CHASM [173] and CanPredict

[174]. Both methods employ a random forest classifier to distinguish driver events from passengers using COSMIC mutations as the positive training set. CHASM trains its model on features such as amino acid substitution properties, evolutionary conservation of the nucleotide position, predicted structure of the mutated position and protein domain annotations, while CanPredict trains on the features such as effect of the mutation on the protein function and Gene Ontology annotations associated with the gene.

These methods aim to detect driver events; however validating the output of these algorithms is very difficult as the only validation is the functional follow up of the event. Thus, it is often common practice to use a combination of these methods to reach a final candidate list of driver events.

Table 3. Computational tools for NGS data analysis

A. Pre-processing tools		
SeqTrim	[175]	
FastX	URL[1]	
ea-utils	URL[2]	
cutadapt	URL[3]	
B. QC programs		
FastQC	URL[4]	
NGS QC Toolkit	[176]	
QC-Chain	[177]	
C. Alignment		
MAQ	[178]	Hash-based short read aligner
BWA, BWA-SW	[179,180]	BWT-based short and long read aligner
ELAND	[181]	Illumina companion hash-based aligner
SSAHA2	[182]	Hash-based short and ling read aligner
Bowtie, Bowtie2	[183],[184]	BWT-based short read aligner
Novoalign	URL[5]	Hash-based short read aligner
SHRiMP, SHRiMP2	[185]	Hash-based short read aligner
Stampy	[186]	Hash-based short read aligner
LAST aligner	[187]	Suffix array based aligner
SOAP2, SOAP3	[188],[189]	Hash-based short read aligner
mrFAST, mrsFAST	[190,191]	Primarily used for detecting structural variants
SARUMAN	[192]	Short read aligner
Corona Lite	Unpublished	Used for SOLiD
BFAST	[193]	Hash-based aligner for SOLiD
BLAT	[194]	Long read aligner
GSNAP	[195]	RNA-seq aligner
STAR	[196]	RNA-seq aligner
MOSAIK	[197]	Hash-based short and long read aligner
FANGS	[198]	Hash-based long read aligner
D. Post-alignment modifications		
Picard	URL[6]	Duplicate removal
GATK	[129]	Indel Realigner
GATK	[129]	Base Quality Score Recalibration

E. Mutation Calling		
SNVMix	[199]	Probabilistic binomial mixture model
SAMTools	[200]	Bayesian SNP calling
GATK Unified Genotyper	[129]	Bayesian SNP calling
SOAPsnp	[125]	Bayesian SNP calling
VarScan/VarScan2	[127,132]	Heuristic variant calling
Atlas-SNP2	[201]	Bayesian SNP calling
F. Small Insertion/Deletion Calling		
Pindel	[133]	
DINDEL	[130]	
VarScan/VarScan2	[127,132]	Heuristic variant calling
SAMTools mpileup	[200]	
G. Copy Number Aberration Detection		
exome2cnv	[202]	
ExomeCNV	[203]	
CoNVEX	[204]	
XHMM	[205]	
CoNIFER	[206]	
H. Gene Expression Quantification and Normalization		
htseq-count	[207]	Gene expression quantification tool from HTSEQ
easyRNASeq	[208]	Gene expression quantification tool from easyRNASeq R package
summarizeOverlaps	[209]	Gene expression quantification tool from GenomicRanges R package
qCount	URL[7]	Gene expression quantification tool from QuasR R package
DESeq	[138]	DESeq normalization and Negative Binomial DE
edgeR	[139]	TMM normalization and Negative Binomial DE
EDASeq	[140]	GC-content normalization
NOISeq	[210]	Non-parametric DE
baySeq	[145]	Negative Binomial DE
BBSeq	[144]	Beta Binomial DE
QuasiSeq	[211]	
EBSeq	[146]	Negative Binomial DE
NBPSeq	[147]	Negative Binomial DE
ShrinkSeq	[212]	Several distributions for DE (incl. Negative Binomial)
SAMSeq	[142]	Non-parametric DE
I. Transcript Assembly and Quantification		
Scripture	[151]	
Cufflinks	[152,153]	
Alexa-seq	[154]	
MISO	[155]	
J. Fusion Detection		
deFuse	[156]	
FusionHunter	[213]	
MapSplice	[157]	
FusionMap	[158]	
FusionFinder	[159]	
Tophat-fusion	[160]	

ChimeraScan	[161]	
K. Somatic Mutation Calling		
MuTect	[162]	Bayesian classifier
VarScan2	[132]	Heuristic variant frequency thresholds
SomaticSniper	[214]	Bayesian model of diploid genotypes based on MAQ
Strelka	[215]	Bayesian model of noisy diploid normal sample and tumour mixture
JointSNVMix	[216]	Bayesian model of diploid genotypes using binomial mixture model
L. Driver Gene Selection		
GenomeMusic	[166]	
MutSigCV	[167]	
CaMP score	[90]	
VariantEffectPredictor	[168]	
SeattleSeq	[169]	
Annovar	[170]	
SIFT	[171]	
PolyPhen	[172]	
CHASM	[173]	
MutationAssessor	[217]	
CanPredict	[174]	

1 http://hannonlab.cshl.edu/fastx_toolkit/

2 https://code.google.com/p/ea-utils/

3 https://code.google.com/p/cutadapt/

4 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

5 http://www.novocraft.com

6 http://picard.sourceforge.net

7 http://www.bioconductor.org/packages/release/bioc/html/QuasR.html

CHAPTER II: RATIONALE AND AIMS

Completion of the human genome sequence and profound advancements in sequencing technologies have enabled fast and systematic analyses of cancer genomes for the identification of genomic aberrations. It is now possible to sequence targeted genomic regions, exomes, genomes and transcriptomes to obtain a comprehensive catalog of mutational profiles of human cancers. The central aim of this thesis is to exploit these technologies to uncover the range of genomic alterations present in the T-cell acute lymphoblastic leukemias (T-ALL) and identify the driver aberrations that contribute to tumorigenesis. A schematic depicting the datasets used in this thesis is presented in Figure 5. For each of the sequencing experiments we have built specific bioinformatics pipelines. Furthermore, we optimized and validated these pipelines (using orthogonal sequencing approaches) in each case individually.

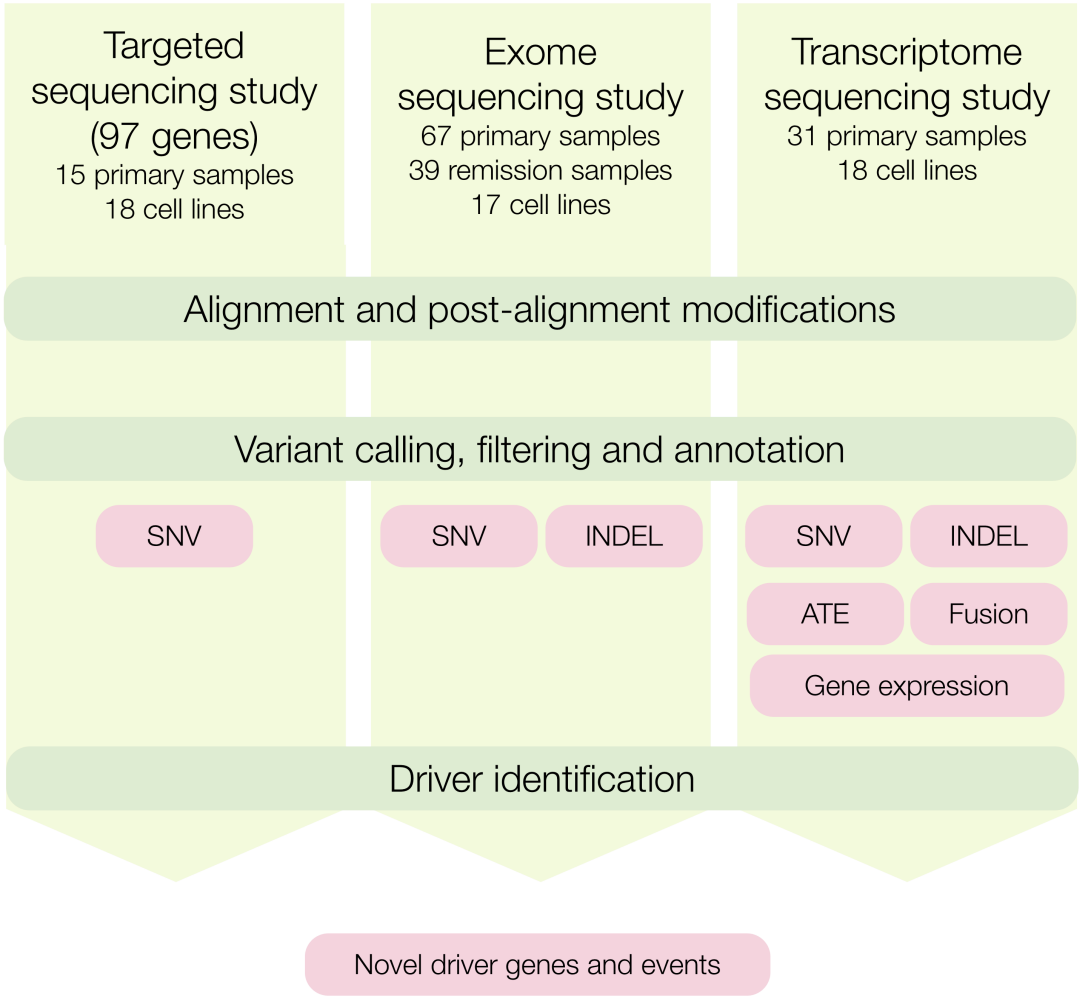
The specific aims of this thesis are:

- Identification of novel driver genes and events involved in the pathogenesis of T-ALL

- Accurate and precise detection of genomic aberrations including SNVs, INDELs, fusions, alternative transcript events as well gene expression levels

- Use of different sequencing applications to decipher the mutational landscape of T-ALL

Figure 5. Schematic overview of the datasets and analysis pipelines used in this thesis



CHAPTER III: RESULTS

PAPER I: High accuracy mutation detection in leukemia on a selected panel of cancer genes

PAPER II: Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia

PAPER III: Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia

Author contributions are detailed in each paper. Overall, the computational experiments and bioinformatics analysis are done by me, while the sequence and functional validation studies are performed in the Laboratory of Molecular Biology of Prof. Jan Cools, Laboratory of Neurogenesis of Prof. Bassem Hassan and Laboratory of Molecular Genetics of Ribosome of Prof. Arlen Johnson.

PAPER I: HIGH ACCURACY MUTATION DETECTION IN LEUKEMIA ON A SELECTED PANEL OF CANCER GENES

Zeynep Kalender Atak[1,*], Kim De Keersmaecker[1,2,*], Valentina Gianfelici[1,2], Ellen Geerdens[1,2], Roel Vandepoel[1,2], Daphnie Pauwels[1,2], Michaël Porcu[1,2], Idoya Lahortiga[1,2], Vanessa Brys[3], Willy G. Dirks[4], Hilmar Quentmeier[4], Jacqueline Cloos[5], Harry Cuppens[3], Anne Uyttebroeck[6], Peter Vandenberghe[1], Jan Cools[1,2], and Stein Aerts [1]

[1] Center for Human Genetics, KU Leuven, Leuven, Belgium.
[2] Center for the Biology of Disease, VIB, Leuven, Belgium.
[3] Genomics Core Facility, University Hospitals Leuven, Leuven, Belgium.
[4] DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Braunschweig, Germany.
[5] Pediatric Oncology/Hematology and Hematology, VU Medical Center, Amsterdam, The Netherlands.
[6] Pediatric Hemato-oncology, University Hospitals Leuven, Leuven, Belgium.

* these authors contributed equally

Published in PLoS One. 2012; 7, e38463

ABSTRACT

With the advent of whole-genome and whole-exome sequencing, high-quality catalogs of recurrently mutated cancer genes are becoming available for many cancer types. Increasing access to sequencing technology, including bench-top sequencers, provide the opportunity to re-sequence a limited set of cancer genes across a patient cohort with limited processing time. Here, we re-sequenced a set of cancer genes in T-cell acute lymphoblastic leukemia (T-ALL) using Nimblegen sequence capture coupled with Roche/454 technology. First, we investigated how a maximal sensitivity and specificity of mutation detection can be achieved through a benchmark study. We tested nine combinations of different mapping and variant-calling methods, varied the variant calling parameters, and compared the predicted mutations with a large independent validation set obtained by capillary re-sequencing. We found that the

combination of two mapping algorithms, namely *BWA-SW* and *SSAHA2*, coupled with the variant calling algorithm *Atlas-SNP2* yields the highest sensitivity (95%) and the highest specificity (93%). Next, we applied this analysis pipeline to identify mutations in a set of 58 cancer genes, in a panel of 18 T-ALL cell lines and 15 T-ALL patient samples. We confirmed mutations in known T-ALL drivers, including *PHF6, NF1, FBXW7, NOTCH1, KRAS, NRAS, PIK3CA,* and *PTEN*. Interestingly, we also found mutations in several cancer genes that had not been linked to T-ALL before, including *JAK3*. Finally, we re-sequenced a small set of 39 candidate genes and identified recurrent mutations in *TET1, SPRY3* and *SPRY4*. In conclusion, we established an optimized analysis pipeline for Roche/454 data that can be applied to accurately detect gene mutations in cancer, which led to the identification of several new candidate T-ALL driver mutations.

INTRODUCTION

Next generation sequencing (NGS) technologies have significantly improved our sequencing capacity in the past five years. They are now widely used for research purposes and are starting to find their way into clinical applications. Although whole genome and whole exome sequencing approaches are successfully implemented for mapping the genomic landscapes of many human diseases, they are not routine strategies for detecting molecular aberrations due to high costs, and long turnover times (run and analysis times). Targeted re-sequencing, on the other hand, is appealing in a clinical setting, given the lower sequencing costs, shorter sequencing time and simpler data analysis. Moreover, as the discovery of novel cancer genes by whole-exome sequencing will gradually saturate and converge into a set of commonly mutated genes in a particular cancer, the identification of these mutations can yield important diagnostic and prognostic information.

Despite the requirement of several days for library preparation and target enrichment for all these platforms, the Roche/454 technology offers the advantages of short run times and data analysis time. In addition, the more restricted data output is also beneficial for turnaround time because fewer patient samples need to be collected to fill an entire sequencing run. Based on these advantages of the 454 platform for sequencing relatively small gene sets, we invested in optimizing bioinformatics pipelines for read mapping and variant calling of 454 reads, with the aim for applying this both for research as well as for clinical purposes. We focused on T cell acute lymphoblastic leukemia (T-ALL), an aggressive hematopoietic cancer caused by malignant transformation of developing T-cells [1]. A set of 97 genes was selected for targeted sequencing. The set consisted of 58 cancer genes [2] and 39 candidate genes including tyrosine kinase and phosphotase coding genes, chromatin modifiers, and several genes belonging to the families of known cancer driver genes such as *TET1-TET3*, or *PIK3CB-PIK3CD-PIK3CG*.

For accurate variant detection, we investigated several existing analysis pipelines and compared their performance. Although the companion software gsMapper is widely

used in the analysis of 454 data [3-5], various alternative mapping and variant calling algorithms have been developed, such as BWA-SW [6] and SSAHA2 [7], BLAT [8] for mapping, and SAMTools [9], VarScan [10], and Atlas-SNP2 [11] for variant calling. Li *et al* [6] reviewed the long read aligners, and Shen *et al* [11] reviewed the variant callers, however, to our knowledge, no comparison has been performed on the combination of mapping and variant calling algorithms in the context of mutation discovery.
Here, we analyzed and compared nine different combinations of a mapping and variant calling algorithms and particularly investigated to what extent low coverage positions can be included in the variation calling process to increase the sensitivity of mutation detection. Next, we apply the optimized pipeline to identify mutations in a set of 58 cancer genes and 39 candidate genes, across 18 T-ALL cell lines and 15 T-ALL patient samples, and identify recurrent mutations in both known and novel drivers.

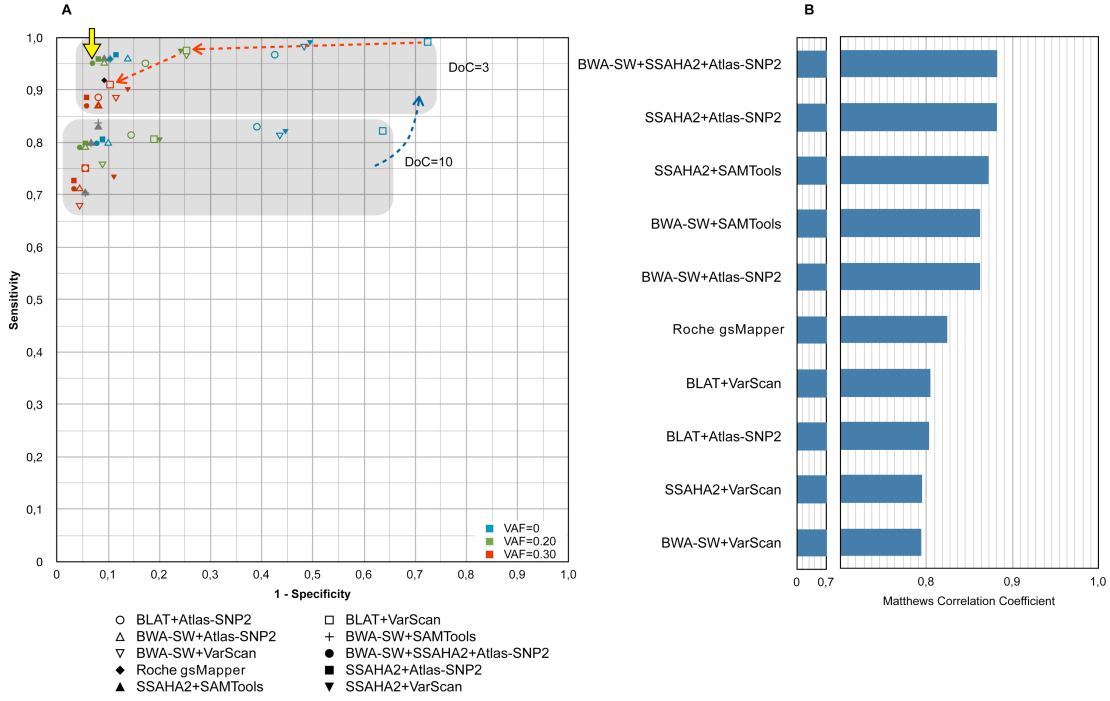
RESULTS

Comparison of mapping and variation calling methods for Roche/454 data

The Roche companion software *gsMapper* is mostly used for the analysis of Roche/454 data. This software first aligns the reads to the reference genome and then lists all positions that are different from the reference genome (variant calling). Although *gsMapper* performed well in several studies [3-5], we wanted to assess its performance on our data set and investigate whether we could achieve better precision and accuracy using alternative aligners and variant callers. We tested eight different combinations of a long read aligner (BWA-SW, SSAHA2, BLAT) and a variant caller (SAMTools, VarScan, Atlas-SNP2) and compared their performance with *gsMapper*.
Each pipeline was applied to the reads obtained from seven T-ALL cell lines and the performance of each pipeline was evaluated by Sanger re-sequencing of 210 candidate variants that were randomly taken from all predicted 8020 variants (containing both SNPs and mutations) from all pipelines. As a measure of the performance of each pipeline, we calculated the Matthews correlation coefficient (MCC), which is a measure of prediction accuracy that is calculated based on the number of successfully predicted true positives and true negatives found by Sanger sequencing (see Materials and Methods). When using default parameter settings **(Table S1)**, the performance of the different pipelines was comparable, with an average MCC of 0.62, with no alternative pipeline performing better than gsMapper (MCC of 0.82) **(Table S1)**.
In NGS studies, the presence of duplicate reads (caused by a PCR amplification step during library preparation) is a potential source of false positive single nucleotide variant (SNV) prediction [12]. Therefore, we added an additional step to remove duplicate reads using Picard, resulting in a 2-24 % increase in MCC, depending on

the pipeline, with an average MCC of 0.73 **(Table S1)**. This showed that duplicate removal is an important step for obtaining correct variant calls.
Next, we further optimized the performance of each pipeline by varying the minimal required number of reads (depth of coverage, DoC) and the minimal required variant reads (variant allele frequency, VAF). Changes in DoC thresholds mainly affected the sensitivity, while varying VAF thresholds affected the predictions in terms of specificity (**Figure 1.A, Table S2**).

Figure 1. Performance comparison and parameter optimization.
(**A**) Different pipelines show different sensitivity and specificity. Varying DoC and VAF thresholds in the variant calling process has an additional affect on the predictions in terms of sensitivity and specificity, respectively. Each pipeline is represented with a different symbol and the performance of each pipeline (in terms of sensitivity and specificity) is plotted under varying DoC and VAF thresholds. Note that the X-axis represents the false positive rate (1-specificity). In this ROC plot, the closer the point to the upper left point of the graph, the better the sensitivity and the specificity. Different colors of the symbols indicate the performance of the pipeline under changing VAF thresholds, and the two shaded boxes indicate the performance under changing DoC thresholds. The plot shows that (i) decreasing the DoC threshold increases the sensitivity of all pipelines as indicated with the blue dotted line; (ii) increasing the VAF threshold increases the specificity with a slight decrease in sensitivity as indicated (in the example of BLAT+VarScan pipeline) with the red dotted line; (iii) the BWA-SW+SSAHA2+Atlas-SNP2 pipeline has the best performance among all pipelines under DoC=3 & VAF=0.20 thresholds as indicated with the yellow arrow. The Roche pipeline is indicated with a black diamond shape since no parameter changes were performed on it, and SSAHA2+SAMTools and BWA-SW+SAMTools pipelines were colored grey since no VAF threshold changes were performed on them. (**B**) The Matthews correlation coefficient for each pipeline is shown for the most optimal performance of that pipeline (**Table S1**). It is interesting to note that the optimal performance of all the pipelines, except Roche gsMapper, was observed for a DoC threshold of 3.



All the pipelines reached their best performance with a DoC threshold of 3, and with a minimum VAF threshold of 0.20 (when applicable) (**Table S1-S2**). In a final effort to minimize false positive predictions, we combined the two best mapping

algorithms in one pipeline, which further increased the sensitivity to 95% and the specificity to 93%. The reason for this increase in accuracy is that certain predicted variants that are caused by erroneous mapping **(Figure S1)** are now filtered out. Although this final pipeline (SSAHA2 + BWA-SW + Atlas-SNP2) performs better than *gsMapper* (91.2% sensitivity and 90.8% specificity), the difference is not large and *gsMapper* can be considered as a valid (and often easy to use) alternative (**Figure 1.B**).

Widespread mutations in cancer genes across 18 T-ALL cell lines and 15 T-ALL patient samples

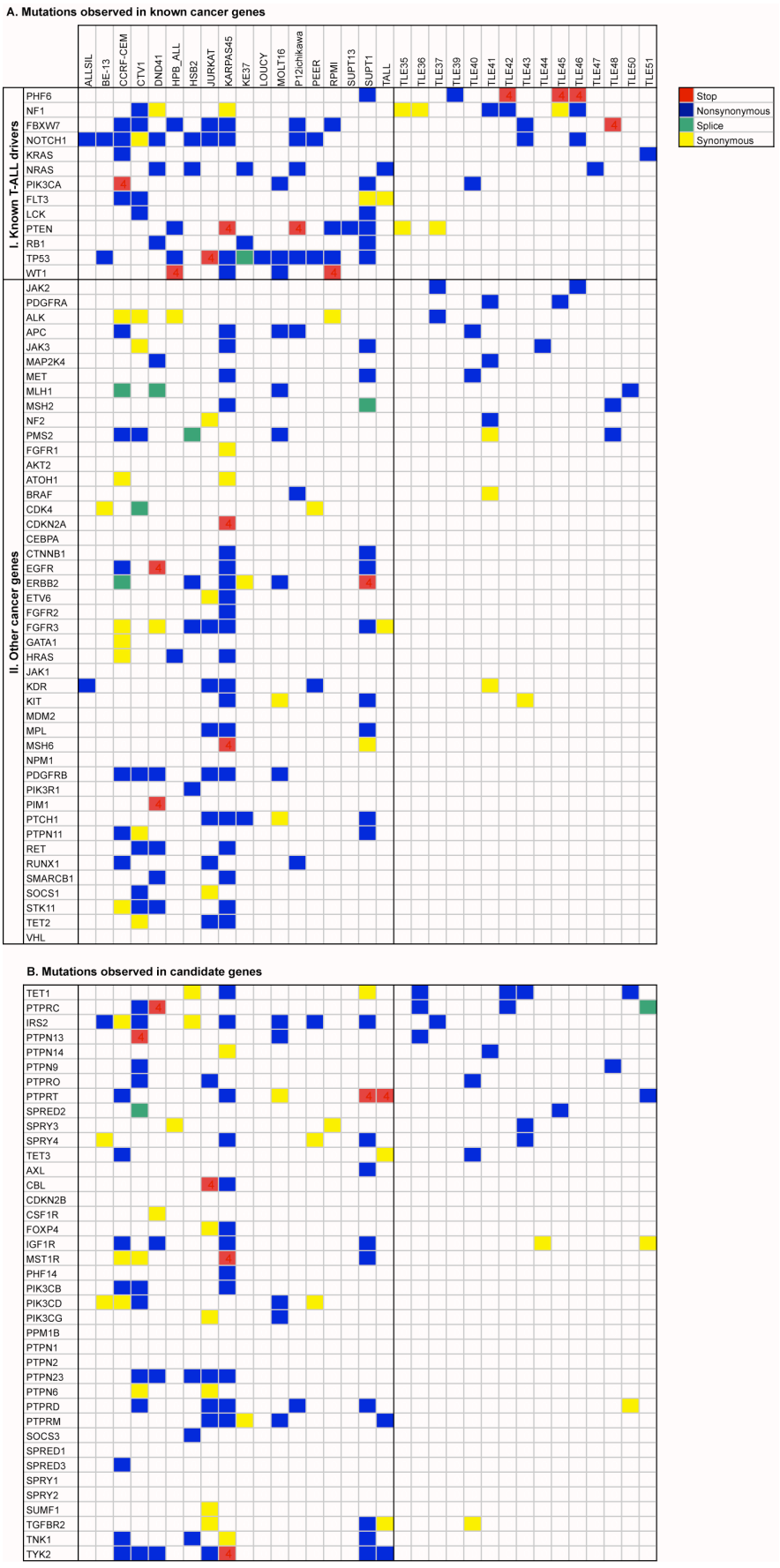
We applied the optimized pipeline determined above, consisting of the SSAHA2+BWA-SW combination for read mapping, and Atlas-SNP2 for variation calling, to identify mutations in a panel of 58 "cancer genes" across 18 T-ALL cell lines and 15 primary T-ALL patient samples. This set of genes consists of 13 T-ALL drivers (**Figure 2.A.I**) and 45 other genes involved in a variety of cancers (**Figure 2.A.II**). All of these genes are present in the Census [2] database of cancer genes except for the recently discovered cancer genes *ATOH1* and *PHF6* [13] [14]. Since *PHF6* mutations are involved in T-ALL we added *PHF6* to our list of T-ALL drivers.

Sequence reads were mapped to the entire reference genome and those reads that map to the selected genes were retained. This resulted in 36% of reads that map to the target sequences on average, with an average coverage of 24.2X and 16.3X for cell lines and patient samples, respectively. Analysis of the sequence data revealed that exons with a very low coverage had a significantly higher GC-content compared to exons with higher coverage (p-value 2.2e-16), a finding consistent with a previously published study [15] **(Figure S2)**. Of the 1565 exons targeted in this study, 18 exons had no coverage in the cell lines or in the patient samples (corresponding to 8710 bps); and 15 exons had no coverage in the patient samples only (corresponding to 5197 bps). On average, 94% and 86% of the targeted exons reached a mean coverage equal or above 3 for the cell lines and the patient samples, respectively.

Variation calling resulted in 836 distinct single nucleotide variants (SNVs) in known cancer genes across the 33 samples. Cell lines had significantly more SNVs in cancer genes than patient samples (p-value <0.001); on average 153 SNVs were detected per cell line and 117 per patient sample. 56% of the predicted SNVs were reported in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) or in the 1000 Genomes project (http://www.1000genomes.org/) and were excluded from further analysis, while the remaining 368 SNVs **(Table S3)** affected 55 of the 58 sequenced cancer genes, primarily in the exons (58.4%) and in untranslated regions (23.9%). Furthermore, there were 8 SNVs affecting splice sites. Of the exonic SNVs, 14 result in the gain of a stop codon (called "stop gain" SNVs), 140 are non-synonymous and the remaining 61 are synonymous coding variations.

To validate the mutations found in cell lines, we compared our results with mutations determined by the Cancer Cell Line project [16], which contains eleven of our 18 cell lines. Of the 35 oncogenic point mutations found in the Cancer Cell line project (determined by capillary sequencing) in the genes that are included in our panel, 31 were recovered by the automated re-sequencing on Roche/454 using the SSAHA2 + BWA-SW + Atlas-SNP2 analysis pipeline, corresponding to a recovery rate of 88.5% **(Table S4)**. Note that gsMapper recovered 30 mutations out of 35, resulting in a recovery rate of 85.7%. The mutations that were missed by Roche/454 sequencing are either due to low coverage at those positions (in two of the four missed mutations, both in *NOTCH1*), or to low variant quality (one TP53 mutation), or to sequencing errors (one *NOTCH1* mutation is covered by 10 reads, none of which contains the variant allele reported by the Cancer Cell line project).
With regards to specificity, both pipelines performed well, for example on the *FBXW7* gene for which we find a protein altering point mutation in exactly the same five cell lines as the Cancer Cell line project (out of the eleven common cell lines). In conclusion, the automated re-sequencing using Roche/454, with either the gsMapper pipeline or the SSAHA2 + BWA-SW + Atlas-SNP2 pipeline, is to a very large extent in agreement with mutations found by capillary sequencing.
Thirteen of the 58 cancer genes have been linked specifically to T-ALL, and we identified protein altering mutations in at least one of these genes in all cell lines and in 10 patient samples (**Figure 2.A.I**). Of the other 45 cancer genes, 36 genes were mutated (**Figure 2.A.II**), of which 25 were mutated in at least two samples (cell line or patient). The genes with most mutations in T-ALL cell lines are *NOTCH1* (non-synonymous mutation in 9/18 cell lines), *TP53* (10/18), *FBXW7* (7/18), and *NRAS* (5/18). These also have mutations in patient samples, except *TP53*, suggesting that it may be easier to obtain cell lines from samples with *TP53* mutation or that *TP53* mutations are acquired during cell culture [17].

Figure 2: Mutations in the 97 genes (next page).
Coding mutations in known cancer genes (**A**) and candidate genes (**B**) are indicated with different color codes. Panel A is further subdivided into (I) genes that are known to be drivers in T-ALL, and (II) the genes that have recurrent somatic mutations in various human cancers. The cell lines are located to the left of the table, and the patient samples are located to the right. Genes are ranked according to the frequency of protein altering mutations in the patient samples.


A. Mutations observed in known cancer genes
ALLSIL
BE-13
CCRF-CEM
CTV1
DND41
HPB_ALL
HSB2
JURKAT
KARPAS45
KE37
LOUCY
MOLT16
P12ichikawa
PEER
RPMI
SUPT13
SUPT1
TALL
TLE35
TLE36
TLE37
TLE39
TLE40
TLE41
TLE42
TLE43
TLE44
TLE45
TLE46
TLE47
TLE48
TLE50
TLE51
I. Known T-ALL drivers
PHF6
NF1
FBXW7
NOTCH1
KRAS
NRAS
PIK3CA
FLT3
LCK
PTEN
RB1
TP53
WT1
II. Other cancer genes
JAK2
PDGFRA
ALK
APC
JAK3
MAP2K4
MET
MLH1
MSH2
NF2
PMS2
FGFR1
AKT2
ATOH1
BRAF
CDK4
CDKN2A
CEBPA
CTNNB1
EGFR
ERBB2
ETV6
FGFR2
FGFR3
GATA1
HRAS
JAK1
KDR
KIT
MDM2
MPL
MSH6
NPM1
PDGFRB
PIK3R1
PIM1
PTCH1
PTPN11
RET
RUNX1
SMARCB1
SOCS1
STK11
TET2
VHL
Stop
Nonsynonymous
Splice
Synonymous
B. Mutations observed in candidate genes
TET1
PTPRC
IRS2
PTPN13
PTPN14
PTPN9
PTPRO
PTPRT
SPRED2
SPRY3
SPRY4
TET3
AXL
CBL
CDKN2B
CSF1R
FOXP4
IGF1R
MST1R
PHF14
PIK3CB
PIK3CD
PIK3CG
PPM1B
PTPN1
PTPN2
PTPN23
PTPN6
PTPRD
PTPRM
SOCS3
SPRED1
SPRED3
SPRY1
SPRY2
SUMF1
TGFBR2
TNK1
TYK2

Identification of recurrent JAK3 mutations in T-ALL

We next determined if mutations in cancer genes could be identified that were previously not linked to T-ALL. We found several such mutations in T-ALL cell lines (**Figure 2.A.II**), but their absence in the patient samples questions their relevance for the pathogenesis of T-ALL.

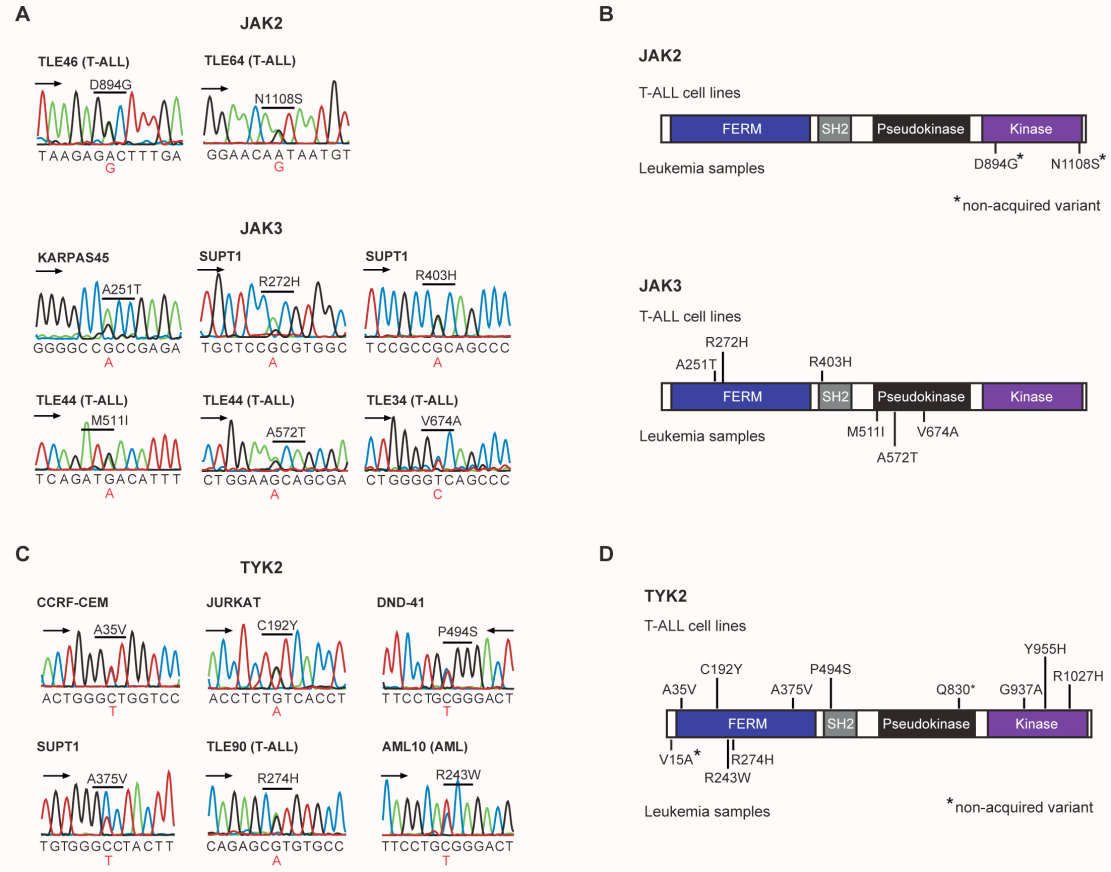
We identified several mutations in *JAK2* and *JAK3* in both cell lines and patient samples. All JAK kinases, except *TYK2* (see below), are known oncogenes in leukemia and activating mutations and translocations affecting *JAK1, JAK2* and *JAK3* were described in multiple, mainly myeloid, hematologic malignancies [18]. Until recently, *JAK1* was the only JAK family member in which point mutations have been described in T-ALL [19]. However, in a recent article *JAK3* gain-of-function mutations were described in T-ALL by Elliott *et al.* [20]. In our study, we have identified 3 non-synonymous coding mutations in 2 patients for *JAK2* (patient TLE37 had two mutations) and 4 non-synonymous coding mutations in 1 patient and 2 cell lines (SUPT1 cell line had two mutations) for *JAK3*. **(Table S3)**. Sanger sequencing confirmed one *JAK2* and all *JAK3* variations **(Table S5, Figure 3.A-B)**. Complementary Sanger sequencing of all exons of the *JAK2* and *JAK3* genes in 31 additional T-ALL patients identified 1 additional *JAK2* variant and 2 additional *JAK3* variants **(Table S5, Figure 3.A-B)**. So, in total, we identified *JAK2* mutations in 2 of 46 (4%) T-ALL samples and in 0 of 18 T-ALL cell lines and *JAK3* mutations in 2 of 46 (4%) T-ALL samples and in 2 of 18 T-ALL cell lines **(Table S5, Figure 3.A-B)**. For *JAK2*, both mutations were also present in a corresponding remission sample, whereas all *JAK3* patient mutations were somatically acquired. Interestingly, patient TLE44 showed 2 somatic mutations in *JAK3*, namely A572T and M511I, which were detected on the same allele (data not shown). Moreover, the M511I mutation has been detected before in AML and over-expression of this mutant transformed IL3 dependent 32D cells and induced T-ALL in mice [21]. Whereas the A572T mutation was not described before, *JAK3* amino acid A572 was found mutated into a V (A572V mutation) in T-cell leukemia, T-cell lymphoma, and AML, and this A572V mutant transformed cytokine dependent hematopoietic cells and induced leukemia in mice [21-24].

Identification of new oncogenes and tumor suppressor genes in T-ALL

Searching for novel T-ALL driver genes can be performed by whole-exome sequencing or other genome-wide approaches. Nevertheless, the Roche/454 platform combined with sequence capture could be useful in a candidate gene approach. In our targeted re-sequencing approach, 39 genes were included that were not causally linked to cancer, but were selected as candidate oncogenes or tumor suppressor genes, because of their function (e.g., tyrosine kinases and tyrosine phosphatases) or because family members had been implicated in cancer (e.g., *TYK2* for the JAK family, *TET1* because *TET2* is a known cancer gene). **Figure 2.B** indicates the

exonic and splice site mutations observed in these genes and the genes were ranked according to the recurrence of protein altering variants across patient samples.

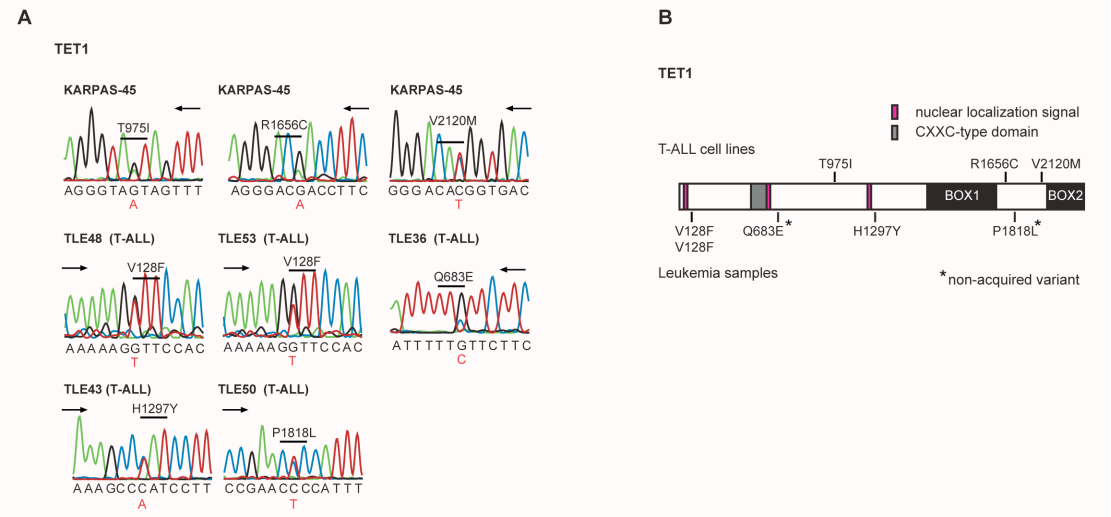
Figure 3: JAK kinase mutations.
(**A**) Sanger sequencing chromatograms corresponding to confirmed *JAK2/JAK3* variants. (**B**) Domain structure of *JAK2* and *JAK3* proteins with indication of novel detected variants. Non-somatic variants are indicated with an asterisk. (**C**) Sanger sequences showing examples of *TYK2* variants detect in T-ALL cell lines or in leukemia patient samples. (**D**) Schematic representation of TYK2 protein structure with indication of all novel *TYK2* variants detected in this study. Non-somatic variants are indicated with an asterisk.



Interestingly, 4 of the 15 sequenced patient samples contain a variation in *TET1*. The TET gene family (*TET1*, *TET2*, *TET3*) of epigenetic regulators is important for the hematology field because of the observation of *TET2* mutations in 10-25% patients with various myeloid hematologic diseases [25-27]. To better assess the mutation frequency of *TET1* in T-ALL, we performed supplemental Sanger sequencing of *TET1* in all cell lines and patient samples and in a panel of 22 additional T-ALL cases. Overall, this resulted in the identification of *TET1* variants in 5/37 (13.5%) of analyzed patients and in 1/18 T-ALL cell lines (KARPAS-45) **(Table S6 and Figure 4)**. The somatic status of detected *TET1* variants was confirmed for 1 case (H1297Y) where a remission sample was available. We also investigated the variants in *TET2* and *TET3* picked up by 454 and performed additional Sanger sequencing for these genes. *TET2* variants were detected in 2 cell lines (JURKAT and KARPAS45) and one *TET3* variant was detected in the

CCRF-CEM cell line, no T-ALL patient samples (0/46) harbored acquired TET2 or TET3 mutations **(Table S6)**.

Figure 4: TET1 mutations in T-ALL.
(**A**) Sanger sequencing chromatograms representing confimed *TET1* variants. (**B**) Schematic representation of TET1 protein structure with indication of all novel *TET1* variants detected in this study. Variants detected in cell lines are depicted above the TET1 protein, variants detected in leukemia patient samples are below the TET1 protein. Non-somatic variants are indicated with an asterisk.

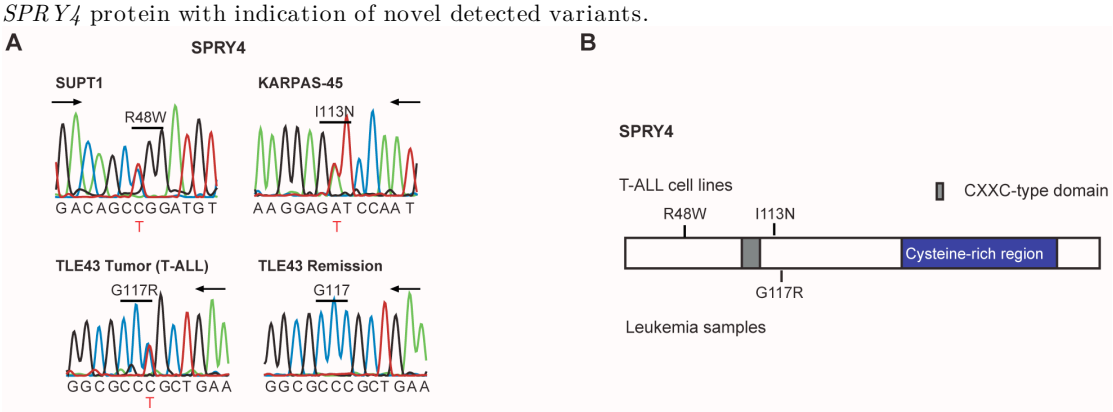


Mutations in tyrosine phosphatase genes, that act as negative regulators of tyrosine signaling, were identified in many T-ALL cell lines and also in several T-ALL patients. Additional mutations in SPRY genes, negative regulators of the RAS/MAPK pathway, were also detected. We identified a homozygous variation in *SPRY3* in one T-ALL patient sample, and 3 mutations in *SPRY4* (2 mutations in cell lines and 1 somatically acquired mutation in a T-ALL patient sample). Sanger sequencing confirmed the presence of these mutations, but did not reveal any additional mutations of *SPRY3/SPRY4* in 22 additional T-ALL cases, bringing the *SPRY4* mutation frequency to 1/37 T-ALL patients and 2/18 T-ALL cell lines **(Table S7, Figure 5)**.

Finally, we also identified several mutations in tyrosine kinases (*IGF1R*, *TYK2*, *TNK1*, and *MST1R*) and associated signaling proteins (*IRS2*, *SOCS3*), but the majority of these mutations were found in cell lines, while primary patient samples showed a much lower frequency of these mutations. The most frequently mutated gene across all cell lines and patient samples was the insulin receptor substrate 2 (*IRS2*) gene, showing non-synonymous coding mutations in 6 cell lines and in one patient sample. Also frequently mutated was *TYK2*, with mutations observed in 6 cell lines; one stop-gain variant and 5 non-synonymous coding variants. Although none of the 15 patient samples carried a mutation in *TYK2*, it could be present at low frequency in patients. To test this, we performed complementary sequencing of *TYK2* in 93 T-ALL, 54 AML and 53 B-ALL patient samples. Despite the high

frequency of *TYK2* variations in T-ALL cell lines, *TYK2* variants were detected only in 2 of 93 T-ALL and 1 of 54 AML cases **(Table S5, Figure 3.C-D)**.

Figure 5: *SPRY4* mutations
(**A**) Sanger sequencing chromatograms showing confirmed *SPRY4* variants. (**B**) Domain structure of the *SPRY4* protein with indication of novel detected variants.



Evidence for the accumulation of specific mutations during in vitro culture of the T-ALL cell lines

The mutation frequency of *TYK2* in T-ALL cell lines compared to primary T-ALL samples was substantially different, with a high mutation rate of *TYK2* in cell lines, but only a low mutation rate in primary samples. To determine if this could be due to the accumulation of *TYK2* mutations during culturing of the cells, we sequenced *TYK2* in different clones of the same T-ALL cell line **(Table 1)**. For the CCRF-CEM cell line, we obtained 5 different subclones that were collected over the years. Interestingly, whereas the R1027H variant was present in all analyzed samples, the A35V variant was only present in our line and in one additional CCRF-CEM clone. In the KARPAS-45 cell line, the Q830* variation was present in 3 different clones. In contrast, only our JURKAT line contained the C192Y mutation, while this was absent in 2 other clones available at DSMZ (www.dsmz.de) **(Table 1)**. These data suggest that at least some *TYK2* mutations were acquired during extended cultivation of the cells, and thus are unlikely to represent an oncogenic event important for the development of leukemia *in vivo*. In addition, analysis of the transforming properties of these mutants in Ba/F3 cells could not identify major differences between wild type *TYK2* and variants of *TYK2* detected in cell lines or patient samples and we could not show any autophosphorylation of *TYK2* in T-ALL cell lines containing *TYK2* variants (data not shown).

These data confirm important differences between cell lines and primary patient samples, which may reflect the accumulation of mutations during *in vitro* cell culture.

Table 1. Analysis of *TYK2* variants in cell lines over time and in different subclones.
Presence of the *TYK2* R1027 and A35V variants was tested in the CCRF-CEM cell line from our group ("CCRF-CEM Cools lab") as well as in the CCRF-CEM cell line as it is currently sold by DSMZ ("CCRF-CEM 2011 DSMZ (ACC240)) and in 5 different CCRF-CEM subclones that DSMZ collected over the years. Similarly, KARPAS-45 from the Cools lab and the KARPAS-45 lines obtained from DSMZ in 2011 and in 1994 were screened for presence of the *TYK2* Q830* variant. JURKAT cells from the Cools lab as well as JURKAT provided by DSMZ in 2011 and 1992 were tested for the *TYK2* C192Y variant.
*: This cell line has 4 copies of chromosome 19 containing *TYK2*. The height of the variant peak on the chromatogram suggests that only 1 copy of *TYK2* contains the Q830* variant

Cell line	Tested variant	Result
CCRF-CEM Cools lab	R1027H	present
CCRF-CEM 2011 DSMZ (ACC240)	R1027H	present
CCRF-CEM subclone 1 DSMZ	R1027H	present
CCRF-CEM subclone 2 DSMZ	R1027H	present
CCRF-CEM subclone 3 DSMZ	R1027H	present
CCRF-CEM subclone 4 DSMZ	R1027H	present
CCRF-CEM subclone 5 DSMZ	R1027H	present
CCRF-CEM Cools lab	A35V	present
CCRF-CEM 2011 DSMZ (ACC 240)	A35V	present
CCRF-CEM subclone 1 DSMZ	A35V	absent
CCRF-CEM subclone 2 DSMZ	A35V	absent
CCRF-CEM subclone 3 DSMZ	A35V	absent
CCRF-CEM subclone 4 DSMZ	A35V	absent
CCRF-CEM subclone 5 DSMZ	A35V	absent
KARPAS-45 Cools lab	Q830*	present*
KARPAS-45 2011 DSMZ (ACC105)	Q830*	present*
KARPAS-45 1994 DSMZ (ACC105)	Q830*	present*
JURKAT Cools lab	C192Y	present
JURKAT 2011 DSMZ (ACC 282)	C192Y	absent
JURKAT 1992 DSMZ (ACC 282)	C192Y	absent

DISCUSSION

We demonstrated that the targeted sequencing approach with an optimized analysis setting can be used to identify oncogenic mutations. This approach could be of particular interest for the detection of point mutations in a set of important oncogenes and tumor suppressors or other disease related genes for diagnosis, prognosis prediction or therapy choice. Such information could be generated in a relatively short timeframe and with unprecedented detail. One of the major advantages over classical Sanger sequencing is the higher throughput of this method allowing that all exons of a gene set of this size can easily be sequenced. As such, full information is provided and rare variants or even previously undiscovered mutations in a particular gene can be detected. Indeed, of the 160 exonic and splice site variants (excluding the 61 synonymous variations) detected in the cell lines and patient samples across our panel of cancer genes, only 40 are found in the COSMIC database [16], of which 24 are associated specifically with T-ALL. Although for some genes mutation hotspots exist (e.g., the *KRAS* G12, G13, Q61 mutations), the function of most cancer genes can be affected by mutations at different positions. Therefore, for most cancer genes the entire coding sequence needs to be re-sequenced, and for this the Roche/454 technology is particularly suitable.

To detect mutations using next-generation sequencing - either to replace or complement molecular diagnosis - standardized bioinformatics analysis pipelines with very high accuracy are required. Such a pipeline consists of a mapping algorithm to align the sequence reads to the reference genome, a variation calling algorithm to identify differences between the sample and the reference, and a variation filtering algorithm.

We compared multiple combinations of mapping and variation calling algorithms, and found that combining two mappers, namely SSAHA-2 and BWA-SW, followed by Atlas-SNP2 yields the most accurate variation detection results. Adding two mapping algorithms filters out false positive variant predictions due to erronous mapping, and the error model of Atlas-SNP2 enables the elimination of reads that have multiple best matches in the reference genome. We also found that additional data filters on depth of coverage and on variant allele frequency further increased both the sensitivity and specifity of variation detection.

We encountered several technical limitations during data analysis. First, we had to remove duplicate reads introduced by PCR amplification steps during sample preparation since we noticed these were causing false positive SNV predicitons. Second, we could only predict SNVs, while indels (small insertions and deletions) had to be ignored since our work (data not shown) and previous studies indicate that 454 reads are not suited for indel detection due to the large amount of false positive results [4]. In a diagnostic setting, where 100% specificity is pursued, it is critical to identify genes or regions in genes that are prone to acquisition of indels and to design alternative assays to investigate them. Likewise, genomic

rearrangements are important causes of T-ALL but require complementary detection technologies.

We believe that using a long read sequencing technology, such as Roche/454 or the more recent Pacific Bioscience, provides particular advantages with regards to both sensitivity and specificity of variation detection. First, long read alignment allows better distinction between highly similar genes in the genome. For example, one of the genes we re-sequenced was *NOTCH1*, a gene with multiple homologs (namely *NOTCH2*, *NOTCH2Nl*, *NOTCH3* and *NOTCH4*). However, we observed no reads mapping to any of these homologs, even though we mapped the reads to the entire genome. This indicates that both the sequence capture and the mapping were specific. On the other hand, we also encountered an example where the sequence capture was not specific. Namely, the *PMS2* gene is one of the targeted genes in our study, yet we observed reads mapping to the *PMS2* pseudogene, *PMS2CL*, which contains the first six exons of *PMS2* gene. Thanks to the use of long reads, this causes no problems for variation detection because for each gene the respective reads mapped *uniquely* to the correct gene, either *PMS2* or *PMS2CL*. Note that the capture technology provides additional cues to achieve higher specificity because not only the exons are covered in the capture but also the flanking intronic regions. Therefore, the alignment is 'aided' by the intronic regions, where sequence similarity between homologs is lower, allowing for the reads to be correctly attributed to their origins in the genome.

Second, mapping long reads to a reference genome is more robust towards extensive local variation, which can be present in particular genomic regions, or can be higher when samples are sequenced from a different ethnicity compared to the reference genome [28]. We indeed found several regions that contain a high number of SNPs within a short sequence window. For example, there are 22 SNP clusters across all samples in a window of 200 bp with at least 3 SNPs, and 5 distinct clusters in a window of 100bp with at least 3 SNPs. Figure S3 shows several examples, such as cluster of three SNPs within 100bp in the *SUMF1* gene in the ALLSIL cell line, and a cluster of 4 SNPs in a 200bp window in the *PTPRM* gene in CCRF-CEM cell line. Nevertheless, in both cases a high coverage is obtained (36x and 46x respectively). These examples show that long reads enable a correct alignment and variation discovery, in contrast to short read sequencing technologies for which the mapping algorithms usually allow for a maximum of two mismathes per read.

We applied our analysis strategy to T-ALL by sequencing a set of 97 genes. This set consists of 58 known oncogenes and tumor suppressors in T-ALL and other cancers, and 39 genes selected via a candidate approach. Regarding the identification of variations in these genes using 454 sequencing and our optimal optimized analysis pipeline, we reached 95% sensitivity and 93% specificity on a confirmation set of 210 variants validated by capillary sequencing. Furthermore, we detected 85.7% of the mutations reported in 11 cell lines that were also sequenced in the Cancer Cell Line project. High performance of our resequencing approach is also illustrated by the fact that we identified mutations in known candidate drivers in T-ALL that were

included in the collection of known cancer genes such as *NOTCH1* [29], *FBXW7* [30], *PTEN* [31], *PHF6* [14], *WT1* [32,33] and *PIK3CA* [34].
We detected mutations in several known cancer genes where a link to T-ALL has not been established yet, such as *JAK3*. Interestingly, a recent article confirmed the mutation status of this gene in the context of T-ALL [20]. We also identified novel mutations in genes that were not previously associated with T-ALL tumorigenesis such as *TET1*, *SPRY3* and *SPRY4*.
It is remarkable that more novel sequence variants are found per cell line sample than per patient, and that genes were in general more frequently mutated in cell lines than in patients. Excessive gene mutations can be explained by potential genomic instability of cells in culture, or can be caused by *in vitro* cell culture conditions. This hypothesis could be confirmed for *TYK2*, a very striking example for which 7/18 (38%) T-ALL cell lines contain novel *TYK2* sequence variants as opposed to only 2/93 (2%) T-ALL patients. Interestingly, we could demonstrate that several *TYK2* variants in cell lines had been acquired during culture. It remains to be determined what is promoting the frequent acquisition of *TYK2* variants in these T-ALL cell lines as opposed to T-ALL patients. The most obvious explanation are differences between the *in vitro* cell culture conditions and the physiological environment of T-ALL cells. As several cytokine signaling pathways depend on TYK2, presence of different cytokines and/or different concentrations of cytokines that use TYK2 signaling might be critical. These observations underscore once more that data obtained from cell culture models should be interpreted with care, especially when extrapolating these data to patient samples.
It is nevertheless interesting to note that this tendency of higher mutation frequency in cell lines compared to patient samples does not extend to all analyzed genes. The most evident example is *TET1*, showing novel variants in only 1/18 cell lines (KARPAS-45) versus 5/37 (13.5%) patients.

In conclusion, we describe a method for fast re-sequencing of a moderate size gene set of 97 genes using 454 next generation sequencing equipment that would be suitable for implementation into the clinic. Our results show that this setting is useful to identify (i) known mutations in known driver genes; (ii) new mutations in known drivers; and (iii) oncogenes or tumor suppressors that had not previously been associated with a specific subtype of cancer based on a candidate gene approach.
The optimized data analysis pipeline, which was assembled from publicly available tools, slightly exceeded the performance of the Roche gsMapper software with 95% sensitivity and 93% specificity for SNV detection, and subsequent analysis of the Roche/454 data from the T-ALL cell lines and patient samples confirmed previously known oncogenes and tumor suppressors in T-ALL and identified previously unrecognized rare somatic mutations in *TET1* and *SPRY4* in T-ALL patients. Screening a larger patient series should reveal the exact mutation frequency of these

genes in T-ALL and whether mutations in these tumor suppressors also play a role in other types of hematopoietic malignancies.

MATERIALS AND METHODS

Targeted genes

97 genes were selected for sequencing in this study. The gene set consists of genes that are known to be involved in oncogenesis of T-ALL (and other cancer types), and a large set of kinases and phosphotases due to their potential therapeutic value. In total, 56 of the selected genes have been causally implicated in cancer according to Census [2] (extracted on 10th of November 2011) and 81 have somatic mutations in cancer according to COSMIC [16] (v48 release). According to the Molecular Signature Database [35] extracted on 10th of September 2010) there are 40 tumor suppressors and 32 oncogenes among the targeted 97 genes. Twenty of the 24 known cancer genes from the NCI-60 cell line set [36] are also included in the selected genes. Functional classification of the genes performed with DAVID [37] shows enrichment for GO terms related to cell proliferation and to signaling cascades, besides the expected enrichment for kinase and phosphotase activity **(Figure S4; Table S8**).

Cell lines and patient samples

All T-ALL cell lines originated from DSMZ (Braunschweig, Germany). Samples from patients with T-ALL (n = 93) , Acute myeloid leukemia (AML) (n = 54) and B-cell acute lymphoblastic leukemia (B-ALL) (n = 53), obtained at diagnosis and remission samples from T-ALL patients (n = 42) were collected at the University Hospital Leuven and VU Medical Center Amsterdam. Diagnosis of T-ALL, AML or B-ALL was based on morphology, cytogenetics and immunophenotyping according to the World Health Organization and European Group for the Immunological Characterization of Leukemias (EGIL) criteria. Informed consent was obtained from all subjects and experiments were approved by the ethical committee of the University Hospital Leuven.

Sequence capture and pyrosequencing

Preparation of a shot-gun DNA sequencing library and capture of the exons, with flanking intron junctions of 97 genes (**Table S9**) was performed on custom designed Nimblegen sequence capture 385K Version 2.0 Arrays (Roche Applied Science, Mannheim, Germany) according to the manufacturer's instructions. The content of these arrays is described in (**Table S9**). Captured DNA was pyrosequenced on a GS FLX instrument (Roche).

Evaluation of the alignment and variant calling algorithms

The performance of the alignment and variant calling algorithms was evaluated to determine the optimal method for analyzing 454 reads. Eight analysis pipelines were constructed from long read aligners BWA-SW [1], SSAHA2 [7], BLAT [8] and variant

callers SAMTools [9], VarScan [10], Atlas-SNP2 [11]. (We will use the term 'pipeline' to refer to a combination of an aligner and a variant caller in the remaining part of this manuscript.) In addition to these pipelines, gsMapper was also included in the evaluation. The aligners map the sequence reads to the human reference sequence (NCBI Build 36.1). To remove duplicate reads in the data, the alignments generated by SSAHA2 and BWA-SW were processed further using Picard [38] (BLAT alignments were not "dedupped" since Picard requires the alignments in BAM format, and format conversion was not possible). Reads mapping to multiple locations in the reference genome (possibly coming from the homologes and/or pseudogenes of the genes targeted in the capture) are marked with a mapping quality of 0.

The pipelines were implemented and reviewed on 7 cell lines: P12ichikawa, KE-37, ALL-SIL, CCRF-CEM, KARPAS45, SUPT1, DND41.

Initial SNV predictions were performed with following settings:

- SAMTools: with pileup -c command, with total coverage threshold of 3 and SNP quality threshold of 20
- VarScan: with pileup2snp –min_coverage 3 –min_reads2 2 min_avg_qual 15 –min_var_freq 0.01 –p_value 0.99
- Atlas-SNP2: with total coverage threshold of 3
- gsMapper: with *HCDiff* (high confidence differences) strategy; requiring the following criteria:
  1. There must be at least 3 reads with the difference.
  2. There must be both forward and reverse reads showing the difference, unless there are at least 5 reads with quality scores over 20 (or 30 if the difference involves a 5-mer or higher).
  3. If the difference is a single-base overcall or undercall, then the reads with the differences must form the consensus of the sequenced reads.

The SNPs to be confirmed with capillary sequencing were selected from the predictions generated with these settings.

Then, predictions from the pipelines were filtered with varying VAF and DoC thresholds. Two VAF thresholds (0.20 and 0.30) and two DoC thresholds (3 and 10) were used. SAMTools pipelines were also processed with *samtools varFilter* command, which implements minimum RMS mapping quality of 25, minimum read depth of 3, maximum read depth of 100, SNP within 10 bp around a gap to be filtered, and maximum number of SNPs in a window of 10 bp to be 2. Atlas-SNP2 pipelines were also filtered with posterior SNP-probability threshold P(SNP|Sj,cj).

The performance of each pipeline was evaluated by Sanger resequencing of 210 variants that were sampled from the pooled set of all predicted variants from all pipelines **(Table S10)** and the performance of each pipeline was quantified by calculating sensitivity, specificity and Matthews correlation coefficient (MCC) which ranges from 0, no correlation, to 1, perfect correlation [39].

Sanger sequencing

Whole genome amplified DNA (REPLI-g system, Qiagen, Hildenberg, Germany) from primary leukemia or remission samples was used as template for PCR amplification of indicated genes. PCR products were Sanger sequenced and inspected for the presence of sequence variants using Mutation Surveyer software (Softgenetics, State College, PA) and CLC DNA Workbench 6 (CLC Bio, Aarhus, Denmark). All variants that were detected in whole genome amplified material were subsequently validated in non-amplified original patient material. Primer sequences are available upon request.

Data availability

Sequence data has been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the EBI, under accession number EGAS00001000268.

Author Contributions

Conceived and designed the experiments: ZKA KDK J. Cools SA. Performed the experiments: ZKA KDK EG VG RV DP MP IL VB HC. Analyzed the data: ZKA KDK EG. Contributed reagents/materials/analysis tools: WGD HQ AU J. Cloos PV. Wrote the paper: ZKA KDK J. Cools SA.

REFERENCES

1. Aifantis, I., Raetz, E. & Buonamici, S. Molecular pathogenesis of T-cell leukaemia and lymphoma. *Nat. Rev. Immunol.* **8,** 380–390 (2008).
2. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4,** 177–183 (2004).
3. Shearer, A. E. *et al.* Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 21104–21109 (2010).
4. Timmermann, B., Kerick, M., Roehr, C. & Fischer, A. Somatic Mutation Profiles of MSI and MSS Colorectal Cancer Identified by Whole Exome Next Generation Sequencing and Bioinformatics Analysis. *PLoS ONE* (2010).
5. Hedges, D. J. *et al.* Exome sequencing of a multigenerational human pedigree. *PLoS ONE* **4,** e8232 (2009).
6. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
7. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res* **11,** 1725–1729 (2001).
8. Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Res* **12,** 656–664 (2002).
9. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
10. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25,** 2283–2285 (2009).
11. Shen, Y. *et al.* A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res* **20,** 273–280 (2010).
12. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491–498 (2011).
13. Bossuyt, W. *et al.* Atonal homolog 1 Is a Tumor Suppressor Gene. *PLoS Biol* **7,** e1000039 (2009).
14. Van Vlierberghe, P. *et al.* PHF6 mutations in T-cell acute lymphoblastic leukemia. *Nat Genet* **42,** 338–342 (2010).
15. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10,** R32 (2009).
16. Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10,** Unit 10.11 (2008).
17. Krieger, D. *et al.* Frequency and clinical relevance of DNA microsatellite alterations of the CDKN2A/B, ATM and p53 gene loci: a comparison between pediatric precursor T-cell lymphoblastic lymphoma and T-cell lymphoblastic leukemia. *Haematologica* **95,** 158–162 (2010).
18. Vainchenker, W., Dusa, A. & Constantinescu, S. N. JAKs in pathology: role of

Janus kinases in hematopoietic malignancies and immunodeficiencies. *Semin. Cell Dev. Biol.* **19,** 385–393 (2008).

19. Flex, E. *et al.* Somatically acquired JAK1 mutations in adult acute lymphoblastic leukemia. *Journal of Experimental Medicine* **205,** 751–758 (2008).
20. Elliott, N. E. *et al.* FERM domain mutations induce gain of function in JAK3 in adult T-cell leukemia/lymphoma. *Blood* **118,** 3911–3921 (2011).
21. Yamashita, Y. *et al.* Array-based genomic resequencing of human leukemia. *Oncogene* **29,** 3723–3731 (2010).
22. Walters, D. K. *et al.* Activating alleles of JAK3 in acute megakaryoblastic leukemia. *Cancer Cell* **10,** 65–75 (2006).
23. Malinge, S. *et al.* Activating mutations in human acute megakaryoblastic leukemia. *Blood* **112,** 4220–4226 (2008).
24. Cornejo, M. G. *et al.* Constitutive JAK3 activation induces lymphoproliferative syndromes in murine bone marrow transplantation models. *Blood* **113,** 2746–2754 (2009).
25. Tefferi, A. *et al.* TET2 mutations and their clinical correlates in polycythemia vera, essential thrombocythemia and myelofibrosis. *Leukemia* **23,** 905–911 (2009).
26. Delhommeau, F. *et al.* Mutation in TET2 in myeloid cancers. *N Engl J Med* **360,** 2289–2301 (2009).
27. Langemeijer, S. M. C. *et al.* Acquired mutations in TET2 are common in myelodysplastic syndromes. *Nat Genet* **41,** 838–842 (2009).
28. Dewey, F. E. *et al.* Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet* **7,** e1002280 (2011).
29. Weng, A. P. *et al.* Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* **306,** 269–271 (2004).
30. Thompson, B. J. *et al.* The SCFFBW7 ubiquitin ligase complex as a tumor suppressor in T cell leukemia. *Journal of Experimental Medicine* **204,** 1825–1835 (2007).
31. Palomero, T. *et al.* Mutational loss of PTEN induces resistance to NOTCH1 inhibition in T-cell leukemia. *Nat Med* **13,** 1203–1210 (2007).
32. Tosello, V. *et al.* WT1 mutations in T-ALL. *Blood* **114,** 1038–1045 (2009).
33. Heesch, S. *et al.* Prognostic implications of mutations and expression of the Wilms tumor 1 (WT1) gene in adult acute T-lymphoblastic leukemia. *Haematologica* **95,** 942–949 (2010).
34. Gutierrez, A. *et al.* High frequency of PTEN, PI3K, and AKT abnormalities in T-cell acute lymphoblastic leukemia. *Blood* **114,** 647–650 (2009).
35. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 15545–15550 (2005).
36. Ikediobi, O. N. *et al.* Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. *Mol. Cancer Ther.* **5,** 2606–2612 (2006).

37. Dennis, G. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4,** P3 (2003).
38. Picard. *picard.sourceforge.net* at <http://picard.sourceforge.net/>
39. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405,** 442–451 (1975).

SUPPLEMENTARY TABLES

Table S1. Performance comparison of different analysis pipelines and parameter settings
Table S2. Performance of the pipelines under varying parameters
Table S3. 368 retained SNVs
Table S4. Point mutations from the Cancer Cell Line project and their detection status by different analysis pipelines
Table S5. Sanger confirmed variants in *JAK* genes
Table S6. Sanger confirmed variants in *TET* genes
Table S7. Sanger confirmed mutations in *SPRY* genes
Table S8. Top 20 enriched (a) molecular function and (b) biological process GO terms from 97 genes
Table S9. 97 selected genes for capture and their targeted exons
Table S10. Selected positions for Sanger sequencing validation

SUPPLEMENTARY FIGURES

Figure S1. Mis-alignment in SSAHA2 causes a false prediction in *TET2*.
Figure S2. Low coverage exons have significantly higher GC-content.
Figure S3. SNP clusters identified in (A) *SUMF1* gene in ALLSIL cell line and (B) *PTPRM* gene in CCRF-CEM cell line
Figure S4. Functional classification of the 97 selected genes based on molecular function terms.

Table S1. Performance comparison of different analysis pipelines and parameter settings

Pipeline		Default parameter settings	Default parameter settings after duplicate removal	Optimized parameter settings after duplicate removal
BWA-SW+SAMTools	MCC	0.68	0.75	0.86
	parameter settings	varFilter & SNP-quality ≥20	varFilter & SNP quality ≥20	DoC ≥3 & SNP quality ≥20
SSAHA2+SAMTools	MCC	0.67	0.74	0.87
	parameter settings	varFilter & SNP-quality ≥20	varFilter & SNP quality≥20	DoC ≥3 & SNP quality ≥20
BWA-SW+VarScan	MCC	0.56	0.66	0.79
	parameter settings	DoC ≥10 & VAF ≥0.25	DoC ≥10 & VAF ≥0.25	DoC ≥3 & VAF ≥0.25
SSAHA2+VarScan	MCC	0.61	0.63	0.79
	parameter settings	DoC ≥10 & VAF ≥0.25	DoC ≥10 & VAF ≥0.25	DoC ≥3 & VAF ≥0.25
BLAT+VarScan	MCC	0.64	0.639*	0.80
	parameter settings	DoC ≥10 & VAF ≥0.25	DoC ≥10 & VAF ≥0.25	DoC ≥3 & VAF ≥0.30
BWA-SW+Atlas-SNP2	MCC	0.59	0.83	0.86
	parameter settings	DoC ≥3 & P(SNP|Sj,cj)$^{\pm}$ ≥0.5	DoC ≥3 & P(SNP|Sj,cj) ≥0.5	DoC ≥3 & P(SNP|Sj,cj) ≥0.5 & VAF ≥0.20
SSAHA2+Atlas-SNP2	MCC	0.63	0.86	0.88
	parameter settings	DoC ≥3 & P(SNP|Sj,cj) ≥0.5	DoC ≥3 & P(SNP|Sj,cj) ≥0.5	DoC≥3 & P(SNP|Sj,cj) ≥0.5 & VAF ≥0.20
gsMapper	MCC	0.82	0.82 §	0.82 §
	parameter settings	HCDiff	HCDiff	HCDiff

MCC = Matthew's Correlation Coefficient

* BLAT pipelines without duplicate removal

§ gsMapper was not subjected to duplicate removal nor further filtering.

± The posterior variant allele probability at locus j when signal is Sj at a specific variant read coverage, c as calculated by Atlas-SNP2.

Table S2. Performance of the pipelines under varying parameters
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Table S3. 368 retained SNVs.
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Table S4. Point mutations from the Cancer Cell Line project and their detection status by different analysis pipelines

Chr	Position (hg19)	Sample	Gene	BWA-SW + SSAHA2 +Atlas-SNP2 pipeline	ROCHE
chr17	7578442	BE-13	TP53	1	1
chr9	139399344	BE-13	NOTCH1	1	1
chr12	25398284	CCRF-CEM	KRAS	1	1
chr13	28608087	CCRF-CEM	FLT3	1	1
chr17	7578406	CCRF-CEM	TP53	1	1
chr17	7577538	CCRF-CEM	TP53	1	1
chr4	153249385	CCRF-CEM	FBXW7	1	1
chr13	28592653	CTV1	FLT3	1	1
chr4	153247289	CTV1	FBXW7	1	1
chr9	139390623	CTV1	NOTCH1	0	0
chr10	89720852	KARPAS45	PTEN	1	1
chr17	7578406	KARPAS45	TP53	1	1
chr17	7577121	KARPAS45	TP53	1	1
chr2	48027431	KARPAS45	MSH6	1	1
chr2	48026605	KARPAS45	MSH6	1	1
chr4	153247289	KARPAS45	FBXW7	1	1
chr9	139397675	KARPAS45	NOTCH1	0	0
chr1	115258747	KE37	NRAS	1	1
chr9	139390816	KE37	NOTCH1	0	0
chr17	7577124	LOUCY	TP53	1	1
chr17	7578457	MOTL16	TP53	1	1
chr1	115258747	P12ichikawa	NRAS	1	1
chr10	89720670	P12ichikawa	PTEN	1	1
chr17	7577538	P12ichikawa	TP53	1	1
chr4	153247289	P12ichikawa	FBXW7	1	1
chr10	89692993	RPMI	PTEN	1	1
chr17	7577121	RPMI	TP53	1	1
chr4	153249384	RPMI	FBXW7	1	1
chr17	7577120	SUPT1	TP53	1	1
chr17	7577138	SUPT1	TP53	1	1
chr17	7577538	SUPT1	TP53	1	1
chr3	178936093	SUPT1	PIK3CA	1	0
chr1	115258747	TALL1	NRAS	1	1
chr17	7577538	TALL1	TP53	0	0
chr17	7577548	TALL1	TP53	1	1

Table S5. Sanger confirmed variants in JAK genes

Sample	JAK gene	nucleotide change (cDNA)	# reads	% mutant reads	AA change	Protein domain	SIFT	Conserved	Somatic	COSMIC
TLE46 (T-ALL)	JAK2	A2681G	21	57%	D894G	Kinase	0,01	yes	no	-
TLE64 (T-ALL)	JAK2	A3323G	N/A	N/A	N1108S	Kinase	0,51	yes	no	33708
KARPAS45	JAK3	G751A	16	69%	A251T	FERM	1	no	N/A	-
SUPT1	JAK3	G815A	18	44%	R272H	FERM	0,16	yes	N/A	-
SUPT1	JAK3	G1208A	8	38%	R403H	SH2	0,13	yes	N/A	-
TLE44 (T-ALL)	JAK3	G1533A	9	44%	M511I	Pseudokinase	0,11	yes	yes	51374
TLE34 (T-ALL)	JAK3	T2021C	N/A	N/A	V674A	Pseudokinase	0,01	yes	yes	-
TLE44 (T-ALL)	JAK3	G1714A	N/A	N/A	A572T	Pseudokinase	0	yes	yes	-
CCRF-CEM	TYK2	G3080A	22	55%	R1027H	Kinase	0	yes	N/A	-
CCRF-CEM	TYK2	C104T	20	60%	A35V	FERM	0,12	no	N/A	-
KARPAS45	TYK2	C2488T	17	35%	Q830*	Pseudokinase	0,76	yes	N/A	-
SUPT1	TYK2	C1124T	18	39%	A375V	FERM	0,2	no	N/A	-
DND41	TYK2	C1480T	13	23%	P494S	SH2	0,86	no	N/A	-
JURKAT	TYK2	G575A	11	55%	C192Y	FERM	0,03	yes	N/A	-
TALL1	TYK2	T2863C	11	36%	Y955H	Kinase	0,07	yes	N/A	-
CTV1	TYK2	G2810C	52	25%	G937A	kinase	0,04	no	N/A	-
TLE10 (T-ALL)	TYK2	T44C	N/A	N/A	V15A	-	0,73	no	no	-
TLE90 (T-ALL)	TYK2	G821A	N/A	N/A	R274H	FERM	0,13	no	N/A	-
AML10 (AML)	TYK2	C727T	N/A	N/A	R243W	FERM	0	no	N/A	-

Variants for which no read # and % mutant reads are reported were identified by supplemental Sanger sequencing. Conservation was assessed by means of the region based conservation score across 46 species (mce46way 30). Somatic status was tested by Sanger sequencing of a sample of the patient at remission, if available. In case a variant has been described in COSMIC before, the COSMIC mutation identifier is reported.

Table S6. Sanger confirmed variants in TET genes

Sample	Gene	Nucleotide change (cDNA)	AA change	# reads	% mutant reads	SIFT	Conserved	Somatic	COSMIC
KARPAS45	TET1	C4966T	R1656C	11	36%	0	yes	N/A	-
KARPAS45	TET1	G6358A	V2120M	N/A	N/A	0,01	yes	N/A	-
KARPAS45	TET1	C2924T	T975I	N/A	N/A	0,19	no	N/A	-
TLE36 (T-ALL)	TET1	C2047G	Q683E	18	44%	0,15	yes	no	-
TLE43 (T-ALL)	TET1	C3889T	H1297Y	21	57%	0,01	yes	yes	-
TLE50 (T-ALL)	TET1	C5453T	P1818L	37	49%	0,02	yes	no	-
TLE48 (T-ALL)	TET1	G382T	V128F	N/A	N/A	0,04	yes	N/A	-
TLE53 (T-ALL)	TET1	G382T	V128F	N/A	N/A	0,04	yes	N/A	-
JURKAT	TET2	G4759A	D1587N	23	48%	0,02	yes	N/A	-
KARPAS45	TET2	A3445G	T1149A	15	47%	0	yes	N/A	-
TLE58	TET2	T2598C	Y867H	N/A	N/A	0,02	yes	N/A	-
TLE44	TET2	G5152T	V1718L	N/A	N/A	0,66	no	no	41742
CCRF-CEM	TET3	A3239C	K1080T	42	40%	0	yes	N/A	-
TLE40	TET3	C1319A	P440H	7	57%	0,01	yes	no	-

Variants for which no read # and % mutant reads are reported were identified by supplemental Sanger sequencing. Conservation was assessed by means of the region based conservation score across 46 species (mce46way 30). Somatic status was tested by Sanger sequencing of a sample of the patient at remission, if available. In case a variant has been described in COSMIC before, the COSMIC mutation identifier is reported.

Table S7. Sanger confirmed mutations in SPRY genes

Sample	Gene	nucleotide change (cDNA)	# reads	% mutant reads	AA change	SIFT	Conserved	Somatically acquired?	COSMIC
TLE43	SPRY3	G785T	22	59	C262F	0	yes	no	-
KARPAS45	SPRY4	T338A	20	50	I113N	0	yes	N/A	-
TLE43	SPRY4	G349A	8	63	G117R	0	yes	yes	-
SUPT1	SPRY4	C142T	7	57	R48W	0	yes	N/A	-

Variants for which no read # and % mutant reads are reported were identified by supplemental Sanger sequencing. Conservation was assessed by means of the region based conservation score across 46 species (mce46way 30). Somatic status was tested by Sanger sequencing of a sample of the patient at remission, if available. In case a variant has been described in COSMIC before, the COSMIC mutation identifier is reported.

Table S8. Top 20 enriched (a) molecular function and (b) biological process GO terms from 97 genes

a) Top 20 biological processes enriched in target genes				
GO_BP	Count	%	P-Value	Corrected p-value (BH)
phosphorus metabolic process	51	52.6	1.3E-32	2.0E-29
phosphate metabolic process	51	52.6	1.3E-32	2.0E-29
enzyme linked receptor protein signaling pathway	31	32.0	2.4E-25	1.9E-22
transmembrane receptor protein tyrosine kinase signaling pathway	27	27.8	4.4E-25	2.3E-22
protein amino acid phosphorylation	36	37.1	3.3E-22	1.3E-19
regulation of cell proliferation	38	39.2	6.5E-22	2.0E-19
phosphorylation	38	39.2	1.1E-21	3.0E-19
regulation of phosphorylation	28	28.9	4.5E-18	1.0E-15
regulation of protein kinase activity	25	25.8	7.0E-18	1.4E-15
regulation of phosphorus metabolic process	28	28.9	1.2E-17	2.2E-15
regulation of phosphate metabolic process	28	28.9	1.2E-17	2.2E-15
regulation of kinase activity	25	25.8	1.5E-17	2.4E-15
protein kinase cascade	25	25.8	3.5E-17	4.9E-15
regulation of transferase activity	25	25.8	3.9E-17	5.1E-15
positive regulation of cell proliferation	23	23.7	5.8E-14	7.0E-12
intracellular signaling cascade	36	37.1	1.6E-13	1.8E-11
protein amino acid dephosphorylation	15	15.5	3.5E-13	3.7E-11
regulation of MAP kinase activity	15	15.5	7.9E-13	7.8E-11
cell surface receptor linked signal transduction	42	43.3	1.3E-12	1.2E-10
dephosphorylation	15	15.5	2.7E-12	2.3E-10
b) Top 20 molecular functions enriched in target genes				
protein tyrosine kinase activity	24	24.7	4.4E-24	1.0E-21
transmembrane receptor protein tyrosine kinase activity	17	17.5	5.8E-21	6.9E-19
protein kinase activity	30	30.9	2.7E-17	2.1E-15
ATP binding	40	41.2	1.3E-14	7.8E-13
adenyl ribonucleotide binding	40	41.2	2.1E-14	9.7E-13
purine nucleoside binding	41	42.3	3.3E-14	1.3E-12
nucleoside binding	41	42.3	4.1E-14	1.4E-12
ribonucleotide binding	43	44.3	1.1E-13	3.2E-12
purine ribonucleotide binding	43	44.3	1.1E-13	3.2E-12
adenyl nucleotide binding	40	41.2	1.1E-13	3.0E-12
protein tyrosine phosphatase activity	14	14.4	2.6E-13	6.0E-12
phosphoprotein phosphatase activity	16	16.5	3.9E-13	8.2E-12
purine nucleotide binding	43	44.3	5.0E-13	9.8E-12
nucleotide binding	43	44.3	9.9E-11	1.8E-9
phosphatase activity	16	16.5	1.5E-10	2.4E-9
kinase binding	14	14.4	2.7E-10	4.3E-9
enzyme binding	20	20.6	2.3E-9	3.3E-8
insulin receptor substrate binding	6	6.2	6.4E-9	8.8E-8
structure-specific DNA binding	11	11.3	5.9E-8	7.7E-7
non-membrane spanning protein tyrosine kinase activity	7	7.2	3.5E-7	4.3E-6

Table S9. 97 selected genes for capture and their targeted exons.
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Table S10. Selected positions for Sanger sequencing validation.
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Figure S1. Mis-alignment in SSAHA2 causes a false prediction in TET2.
IGV (Integrative Genomic Viewer) software visualizes alignment of next generation sequencing reads to the reference sequence.15 In IGV, each sequence read is represented as a grey rectangle and the reference sequence is represented at the bottom. If there is base in a read that is different from the reference sequence, it is indicated with the corresponding letter. This figure shows the IGV output when anlayzing the same set of reads with the BWA-SW (top) and with the SSAHA2 (bottom) algorithms for sequence alignment. Looking at the alignments generated by these two algorithms revealed that SSAHA2 alignment was incorrectly positioning 3 reads (as indicated by colored read names on the plot), causing a false variant call on chr4:106384366 and resulting false prediction of a non-synonymous coding mutation in the *TET2* gene.

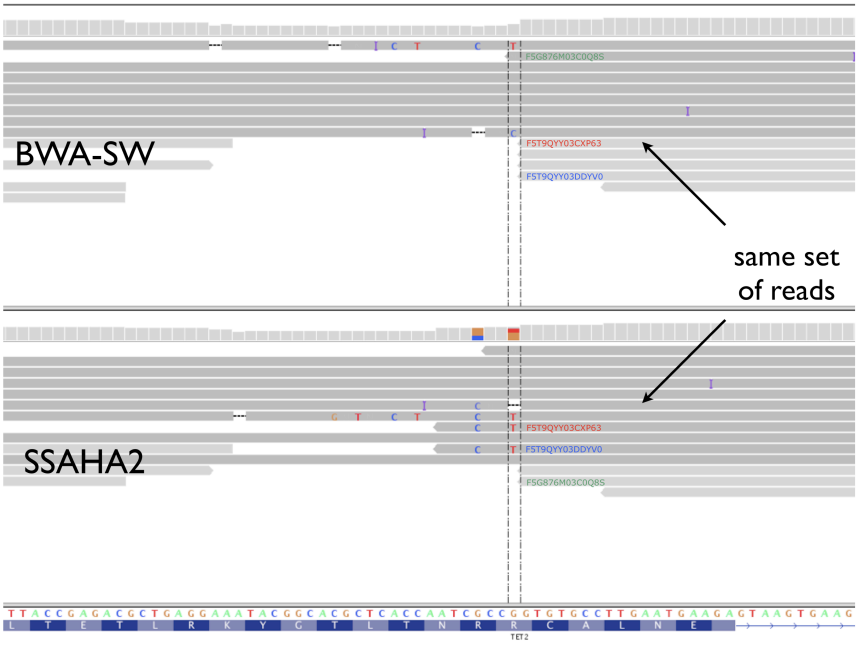


Figure S2. Low coverage exons have significantly higher GC-content.
Comparing the GC-content of the exons with low and high coverage revealed that the two groups have significantly different GC-content (p-value 2.2e-16), with low coverage exons having higher GC-content.

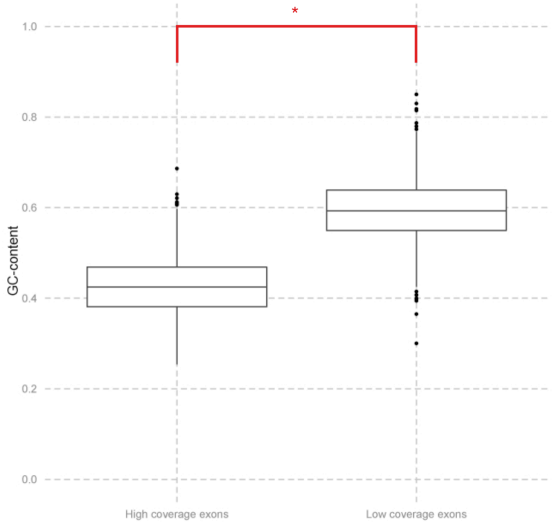


Figure S3: SNP clusters identified in (A) *SUMF1* gene in ALLSIL cell line and (B) *PTPRM* gene in CCRF-CEM cell line

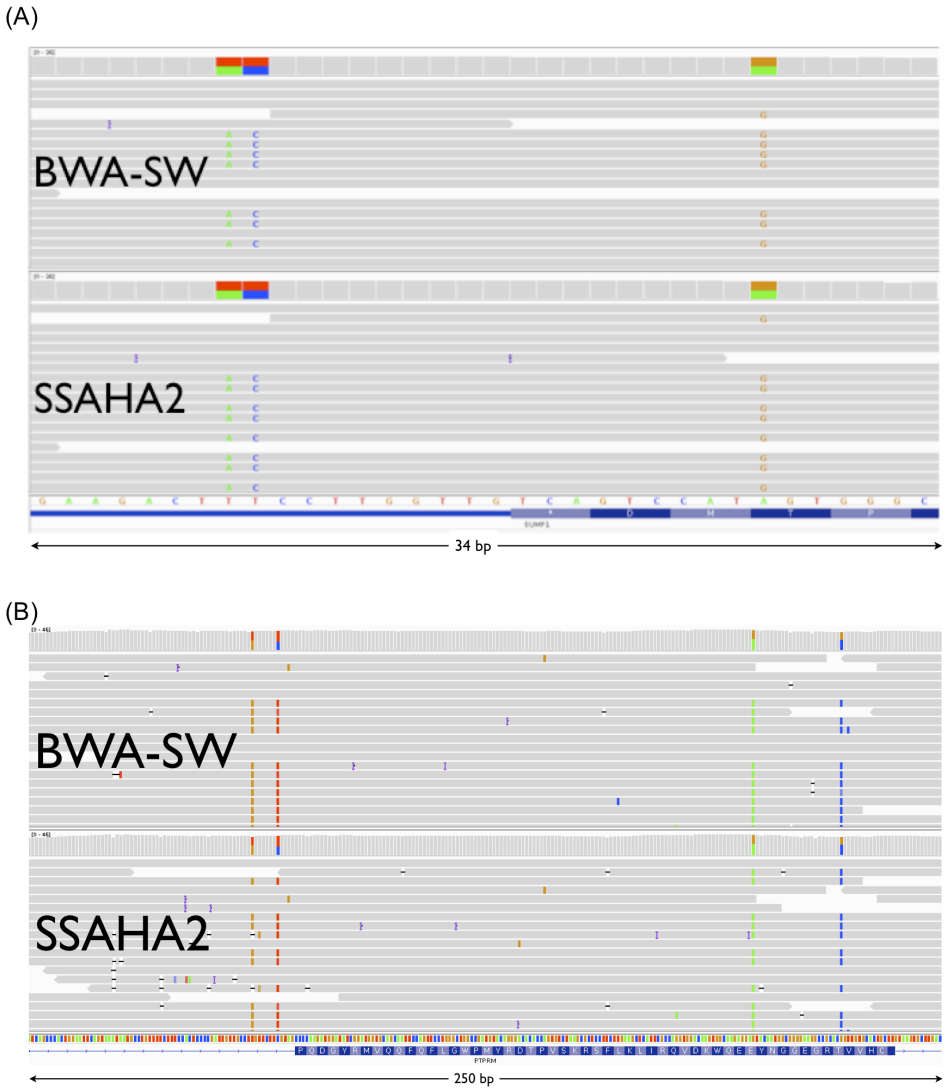
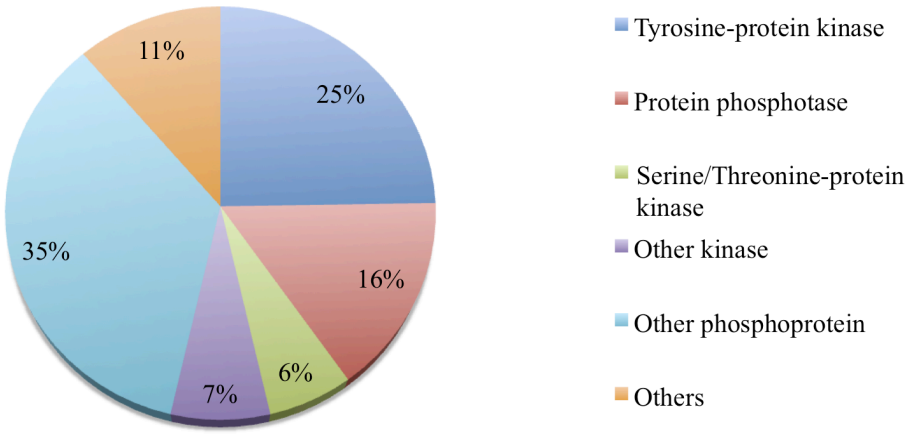


Figure S4: Functional classification of the 97 selected genes based on molecular function terms.



PAPER II: EXOME SEQUENCING IDENTIFIES MUTATION IN CNOT3 AND RIBOSOMAL GENES *RPL5* AND RPL10 IN T-CELL ACUTE LYMPHOBLASTIC LEUKEMIA

Kim De Keersmaecker[1,2,9], Zeynep Kalender Atak[1,9], Ning Li[3,9], Carmen Vicente[1,2], Stephanie Patchett[4], Tiziana Girardi[1,2], Valentina Gianfelici[1,2], Ellen Geerdens[1,2], Emmanuelle Clappier[5], Michaël Porcu[1,2], Idoya Lahortiga[1,2], Rossella Lucà[1,2], Jiekun Yan[1,2], Gert Hulselmans[1], Roel Vandepoel[1,2], Bram Sweron[1,2], Kris Jacobs[1,2], Nicole Mentens[1,2], Iwona Wlodarska[1], Barbara Cauwelier[6], Jacqueline Cloos[7], Jean Soulier[5], Anne Uyttebroeck[8], Claudia Bagni[1,2], Bassem A. Hassan[1,2], Peter Vandenberghe[1], Arlen W. Johnson[4], Stein Aerts[1,9], Jan Cools[1,2,9]

[1] Center for Human Genetics, KU Leuven, Leuven, Belgium
[2] Center for the Biology of Disease, VIB, Leuven, Belgium
[3] BGI Europe, Copenhagen, Denmark
[4] Section of Molecular Genetics and Microbiology, Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, TX, USA
[5] Hôpital Saint-Louis, Paris, France
[6] AZ St-Jan, Brugge, Belgium
[7] Pediatric Oncology/Hematology and Hematology, VU Medical Center, Amsterdam, The Netherlands.
[8] Pediatric Hemato-Oncology, University Hospitals Leuven, Leuven, Belgium
[9] equal contribution

Published in Nature Genetics. 2013; 45,186-190.

ABSTRACT

T-cell acute lymphoblastic leukemia (T-ALL) is caused by cooperation of multiple oncogenic lesions[1,2]. We used exome sequencing on 67 T-ALLs to gain insight in the mutational spectrum in these leukemias. We detected protein-altering mutations in 508 genes, with an average of 8.2 mutations in pediatric and 21.0 in adult T-ALL. Using stringent filtering, we predict 7 novel oncogenic driver genes in T-ALL. We identify *CNOT3* as a tumor suppressor mutated in 7/89 (7.9%) of adult T-ALL

and for which knock-down causes tumors in a sensitized drosophila model[3]. In addition, we identify mutations in the ribosomal proteins RPL5 and RPL10 in 12/122 (9.8%) of pediatric T-ALL, with recurrent mutation of arginine 98 in RPL10. Yeast and lymphoid cells expressing the RPL10 p.Arg98Ser mutant showed a ribosome biogenesis defect. Our data provide insights in the mutational landscape of pediatric versus adult T-ALL and identify the ribosome as a potential oncogenic factor

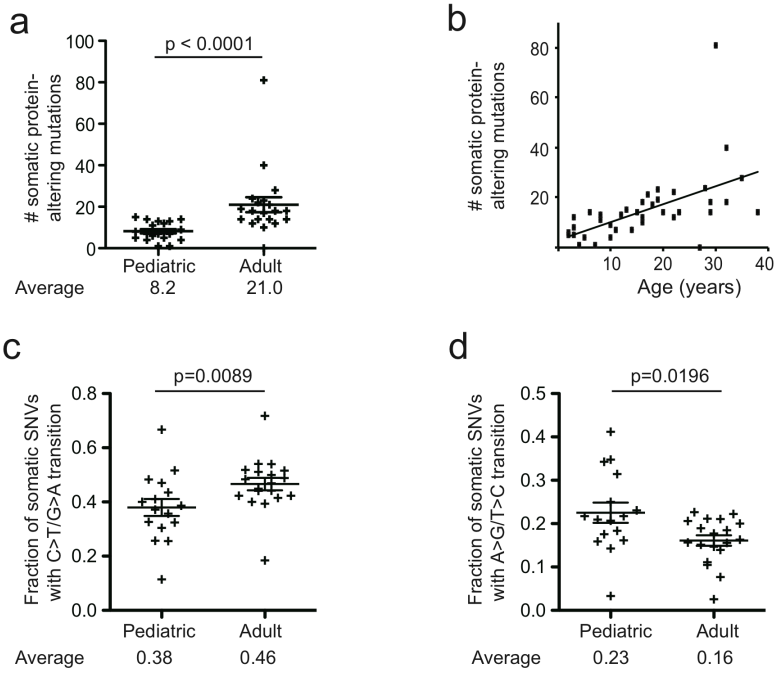
RESULTS AND DISCUSSION

T-ALL is a genetically heterogeneous leukemia that is caused by accumulation of multiple oncogenic lesions, which have been identified through characterization of chromosomal aberrations or via candidate gene sequencing[4-7]. In addition, recent whole genome sequencing of 12 immature early T-cell precursor (ETP) ALLs revealed several new oncogenic drivers in this T-ALL subtype[8]. To discover novel disease driving genes in pediatric and adult T-ALL, we performed exome sequencing on 67 diagnostic T-ALL samples, 39 corresponding remission samples and 17 cell lines (**Supplementary Tables 1-3**).

For discovering somatic mutations, we limited our initial analysis to the 39 paired diagnosis-remission samples. To assess the performance of variant calling, 185 predicted single nucleotide variations (SNVs) were validated by Sanger sequencing. This set was used to detect the filtering strategy with the best sensitivity-specificity characteristics. Different parameters were tested as filters, including coverage of the variant nucleotide, variant allele frequency, variant quality, and presence in repeat regions. Finally, ranging thresholds of the "somatic score", as calculated by SomaticSniper[9] were used as filter (**Supplementary Fig. 1**). Removing variants with a somatic score below 70 resulted in 89% sensitivity and 96% specificity. Using this filter, a second batch of SNVs was selected for testing with capillary sequencing, which confirmed 80% (67 out of 84) of predicted SNVs.

We identified 1810 somatic SNVs and 1248 insertion-deletions (INDELs) in the 39 diagnostic-remission pairs. Excessively high numbers of somatic INDELs were present in 3 samples, possibly due to defective DNA repair. These INDELs were excluded for candidate gene detection. One fourth of the somatic mutations were protein-altering, with the majority being missense mutations (413), and the rest frame-shift INDEL (55), in-frame INDEL (30), nonsense coding (32) or splice site mutations (39). On average, each sample contained 14.7 somatic protein altering SNVs and INDELs (**Supplementary Table 4**). Remarkably, adults (age >15 years) showed 2.5 times more somatic protein-altering mutations than children (21.0 versus 8.2; $p<0.0001$) and there was a clear correlation between patient age and mutation number (**Fig. 1a-b, Supplementary Fig. 2**). However, outcome was not linked to mutation number (**Supplementary Fig. 3**). Interestingly, a larger fraction of somatic SNVs in adults were C>T transitions and adults had a lower fraction of A>G transitions than children **(Fig. 1c-d)**.

Figure 1. Correlation between patient age and mutation number and type.
(**a**) Plot showing the number of protein-altering somatic mutations in pediatric (≤ 15 years) and in adult (≥ 16) T-ALL patients. Average and s.e.m. is indicated on the plots. The p-value tests whether there is a significantly different mutation number in adults versus children and was calculated using a 2-tailed Wilcoxon signed rank test. Group size pediatric: n=19; adult: n=20. (**b**) Dot plot representing the number of protein-altering somatic mutations versus patient age. (**c, d**) Plot showing the fraction of somatic SNVs that were C>T/G>A transitions (**c**) or A>G/T>C transitions (**d**) in pediatric and in adult T-ALL patients. Average and s.e.m. is indicated on the plots. Samples with less than 10 somatic SNVs were excluded for this analysis. The reported p-value tests whether there is a significant difference between adults and children and was calculated using a 2-tailed Wilcoxon signed rank test. Group size pediatric: n=16; adult: n=19.



Protein-altering mutations occurred across 508 genes **(Supplementary table 5)**. To distinguish driver from passenger mutations, we only considered genes that were mutated in at least 2 samples and that were significantly more mutated than the local background mutation rate as calculated by Genome MuSiC[10] **(Supplementary table 6)**. We identified 15 candidate drivers meeting these two criteria (**Table 1, Fig. 2**), and 11 additional genes that were recurrently but not significantly mutated **(Supplementary Fig. 4)**. Of the 15 candidate drivers, 8 were known drivers in T-ALL and 7 were novel. Reassuringly, we found additional mutations in many of the 15 candidate drivers across the 28 additional diagnosis samples and 17 cell lines that were sequenced (**Fig. 2**, **Supplementary Fig. 5, Supplementary table 7**).

Adult samples showed 2.7 times more mutations in candidate drivers than children (1.9 versus 0.7; p=0.0034) (**Supplementary Fig. 2**). Moreover, mutations in *FBXW7*, *CNOT3*, *PHF6*, *KDM6A* and *MAGEC3* were mainly in adults whereas *RPL10* mutations were almost exclusively found in children (**Fig. 2+3, Table 1,**

Supplementary tables 8-10). Notably, our candidate driver list included *RPL5* and *RPL10,* two genes encoding ribosomal proteins that occupy neighboring positions in the 60S ribosomal complex **(Supplementary Fig. 6)**, with 5 exome samples carrying the same somatic p.Arg98Ser mutation in RPL10. Also the *CNOT3* gene showed a mutational hotspot, with 3 patients carrying a p.Arg57 substitution **(Fig. 3a)**. Mutation screening of these 3 genes in an independent confirmation cohort of 144 T-ALLs identified additional mutations in each of these genes **(Fig. 3a, Supplementary table 8-9)**, resulting in total mutation frequencies of 8/211 (3.8%) for *CNOT3* and 15/211 (7.1%) for *RPL5* and *RPL10*. Adding the results from the confirmation cohort consolidated the association between *CNOT3* mutations and adult age (p= 0.01) and *CNOT3* was mutated in 7/89 (7.9%) of adult T-ALLs. In contrast, *RPL10* mutations were associated with young age (p= 0.03) and 10/122 (8.2%) pediatric cases showed *RPL10* mutations **(Fig. 3b)**. Mutations in *CNOT3*, *RPL10* or *RPL5* were not associated with any of the major molecular subgroups in T-ALL, nor with *NOTCH1* mutations **(Supplementary Table 8-10)**.

Ribosomal defects have been identified in inherited hematopoietic disorders ('ribosomopathies') that result in anemia and a propensity to develop leukemia[11]. Mutations in *RPL5* have previously been associated with Diamond Blackfan anemia and were studied in much detail in that context[11], but mutation of *RPL10* has not been described in any disease. Interestingly, also loss of RPL22, another 60S ribosomal protein, was recently identified in T-ALL[12], and also in our exome cohort we detected 1 patient with an RPL22 frameshift mutation **(Supplementary Table 5)**. *RPL10* is located on the X chromosome, and 7/11 mutant cases were males carrying the mutation in nearly all leukemia cells. Moreover, the single *RPL10* mutated female from whom we had RNA available expressed only the mutant allele in the tumor cells **(Supplementary Figure 7)**. To confirm that these *RPL10* mutations were not random passenger mutations but alter RPL10 function, we engineered yeast cells to express Rpl10 wild type, Rpl10 p.Arg98Ser, Rpl10 p.Arg98Cys or Rpl10 p.His123Pro as sole copy. Rpl10 has been intensively studied[13] and is highly conserved in yeast, with Arg98 being unchanged from yeast to human (**Supplementary Fig. 8a)**. Interestingly, residues Arg98 and His123 are closely apposed in a beta-hairpin near the peptidyltransferase center, the catalytic core of the ribosome **(Supplementary Fig. 8b-c)**.

Figure 2. Overview of mutations in 15 identified candidate T-ALL driver genes in 67 patient samples. (next page)

Mutations in 15 candidate T-ALL driver genes across the patient set. For clarity, only patients harboring mutations in any of these 15 genes are shown. Each type of mutation is indicated with a different color as indicated in the legend and symbols for homozygous, hemizygous, and compound heterozygous mutations are explained. Mutations with no indication are heterozygous. All mutations in this figure were validated by Sanger sequencing. Relevant patient characteristics (identified by Sanger sequencing, karyotyping, or gene expression) are included at the bottom of the figure. Mutations in *NOTCH1* were hard to identify by exome sequencing due to low capture efficiency and resulting low sequence coverage of *NOTCH1*. *NOTCH1* mutations detected by Sanger sequencing are indicated in the section patient characteristics of the figure. Detailed description of the mutations shown in this figure is in supplementary tables 5, 7, 8 and 9.


Somatic splice site
Somatic missense
Somatic nonsense
Somatic frame-shift
Somatic in-frame
Homozygous
Hemizygous
Splice site
Missense
Nonsense
Frame-shift
In-frame
Compound heterozygous
Wild-type
Mutated
Deleted
Duplicated
Fusion
Overexpressed
Male
Adult
Pediatric
Transcriptional regulation
Translational regulation
Signalling & Adhesion
Epigenetic regulation
Other
Patient Characteristics
NOTCH1
FBXW7
WT1
BCL11B
CNOT3
RPL10
RPL5
JAK3
PTEN
DNM2
ODZ2
PHF6
TET1
KDM6A
MAGEC3
NOTCH1 (PEST)
NOTCH1 (HD)
PTPN2 status
MYB status
IL7R status
ABL1 status
MLL status
TLX3 expression
TLX1 expression
TAL1 expression
Gender
Age

Table 1. Significantly and recurrently mutated genes

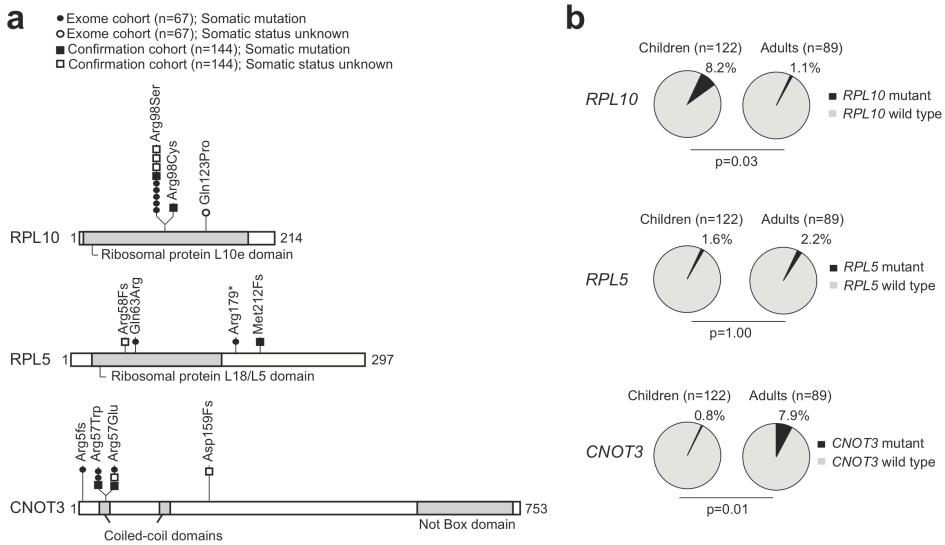
Gene	# mutant patients	Gene function	Associated age group	Associated pathologies with genomic alterations
NOTCH1	29/67 (43.3%)*	transmembrane receptor, releases intracellular NOTCH1 transcriptional enhancer upon acitvation	none	T-ALL[25]; CLL[26-27]; lung cancer[28]; head and neck cancer[29-30]; breast cancer[31-32]
FBXW7	8/67 (11.9%)	Part of ubiquitin-ligase complex targeting cyclin E, MYC and NOTCH1	none	T-ALL[33]; various cancer types[34]
WT1	7/67 (10.4%)	zinc finger transcription factor	none	T-ALL[35]; AML[36]; Wilms tumor
BCL11B	5/67 (7.5%)	zinc finger transcription factor	none	T-ALL[37]
CNOT3	8/211 (3.8%)	part of the CCR4-NOT complex that regulates gene expression	adult	
RPL10	11/211 (5.2%)	ribosomal protein of the 60S ribosomal subunit	pediatric	autism[38-39]
RPL5	4/211 (1.9%)	ribosomal protein of the 60S ribosomal subunit	none	Diamond Blackfan anemia[11]
JAK3	7/67 (10.4%)	kinase involved in cytokine receptor signaling	none	T-ALL[8,40]; various myeloid and lymphoid malignancies
PTEN	4/67 (6.0%)	phosphatase antagonizing PI3K function	none	T-ALL[18]; various cancer types
DNM2	4/67 (6.0%)	microtubule associated GTPase	none	T-ALL[8]; Charcot Marie Tooth disease[41]; centronuclear myopathy[41]
ODZ2	2/67 (3.0%)	may function as a cellular signal transducer	none	
PHF6	12/67 (17.9%)	plant homeodomain-like finger (PHF) family protein that may regulate transcription	adult	T-ALL[42]; AML[43]; Borjeson-Forssman-Lehmann syndrome[44]
TET1	4/67	epigenetic regulator	none	t(10;11)(q22;23) (MLL-

	(6.0%)	converting methylcytosine (5mC) to 5-hydroxymethylcytosine (hmC)		TET1 fusion) in AML and B-ALL[45,46,47]
KDM6A	3/67 (4.5%)	histone demethylase for 'Lys-27' of histone H3	none	various cancer types[48-49]; Kabuki syndrome[50]
MAGEC 3	2/67 (3.0%)	Gene function unknown. Only expressed in normal testis and in various tumor types.	none	Diffuse large B-cell lymphoma[51]

*Note: Mutations in *NOTCH1* were hard to identify by exome sequencing due to poor capture efficiency and resulting low sequence coverage of *NOTCH1*. The reported mutation number in this table reflects *NOTCH1* mutations detected by complementary Sanger sequencing.

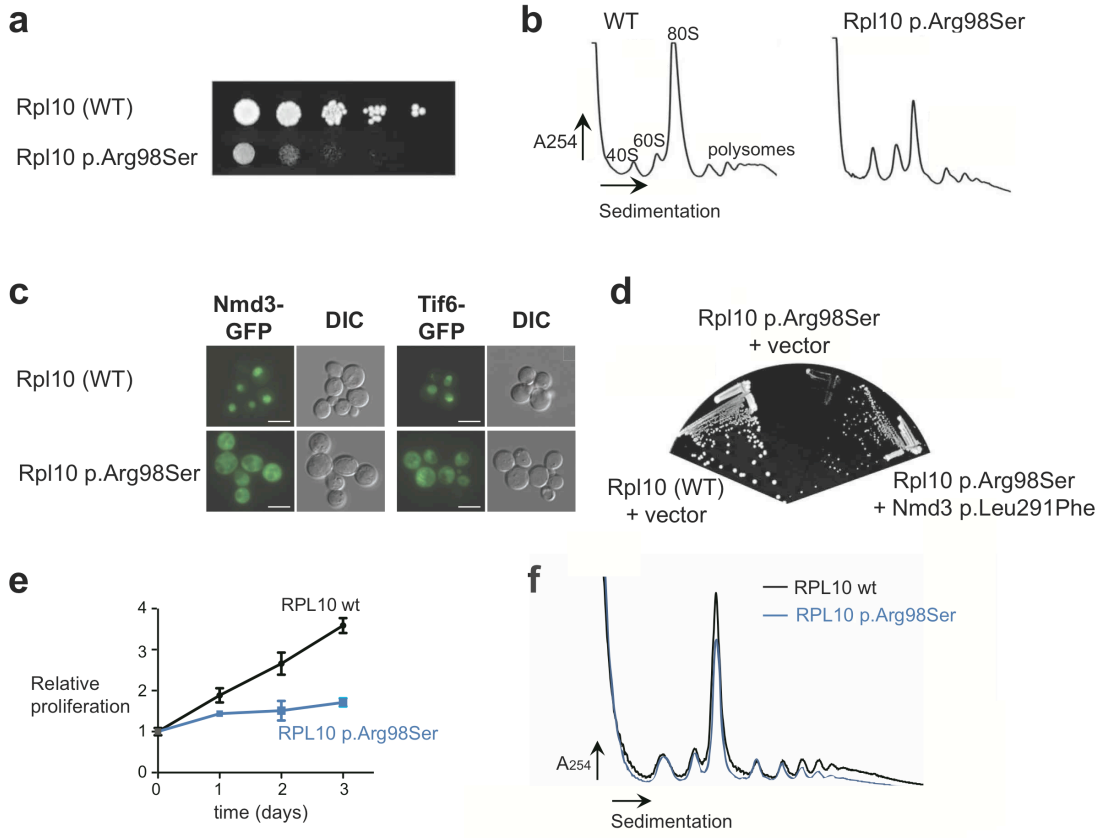
In yeast, expression of the Rpl10 mutants impaired proliferation and caused a ribosome biogenesis defect, evidenced by the altered ratio of mature 80S and free subunits and reduced presence of polysomes (**Fig. 4a-b, Supplementary Figure 9**). In addition, Nmd3 and Tif6 showed aberrant accumulation in the cytoplasm in cells expressing Rpl10 p.Arg98Ser (**Fig. 4c, Supplementary Figure 9**), demonstrating that this mutation impaired release of the 60S export adapter Nmd3 as well as the subunit anti-association factor Tif6.

Figure 3. Overview of mutations in RPL10, RPL5 and CNOT3.
(**a**) Schematic representation of RPL10, RPL5 and CNOT3 protein structure with indication of the mutations detected in 211 T-ALL samples. Somatic status of the mutations is indicated as explained in the figure legend. Supplementary tables 8 and 9 report on the characteristics of the patients with *RPL10*, *RPL5* or *CNOT3* mutations. (**b**) Pie diagrams reporting mutation frequencies detected in adult versus pediatric patients. All reported p-values test whether there is a significant difference between mutation frequency in adults versus children and were calculated using the unpaired t-test.



The deleterious effects of the Rpl10 mutants were partially suppressed by Nmd3 p.Leu291Phe (**Fig. 4d**), a mutant with weakened affinity for the ribosome[14] and by increased gene dosage for Nmd3 (**Supplementary Figure 9**).These data indicate that these Rpl10 mutants affect release of Nmd3 from the ribosome. Retention of Nmd3 and Tif6 on pre-60S subunits blocks ribosome assembly and the resulting depletion of Nmd3 from the nucleus reduces export of new ribosome subunits[15]. We also tested the effect of expression of human RPL10 p.Arg98Ser, the most frequent RPL10 mutation, in lymphoid cells. Also in these cells, expression of RPL10 p.Arg98Ser resulted in a proliferation and ribosome biogenesis defect (**Fig. 4e-f**).

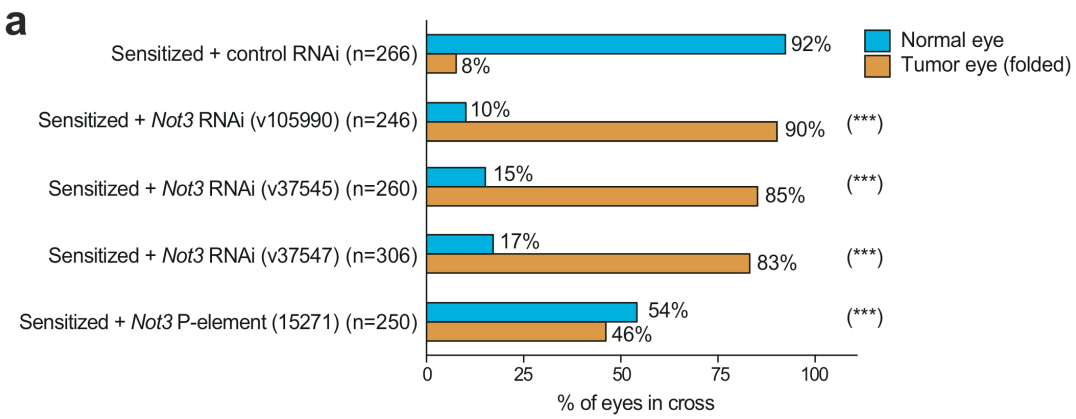
Figure 4. Cellular effects of RPL10 p.Arg98Ser mutation.
The growth of yeast cells expressing wild type (WT) Rpl10 or Rpl10 p.Arg98Ser was compared by plating ten-fold serial dilutions (**a**), and polysome profiles were obtained (**b**). Fluorescence of Nmd3-GFP and Tif6-GFP was examined in WT and Rpl10 p.Arg98Ser-expressing cells. Scale bars: 5 μm. (**c**). In the case of Nmd3, cells also contained a leptomycin B (LMB)-sensitive Crm1 and Nmd3-GFP localization was examined after treatment with LMB to trap Nmd3 in the nucleus. (**d**) Rpl10 WT or p.Arg98Ser yeast cells were transformed with vector or vector expressing Nmd3 p.Leu291Phe. Ten-fold serial dilutions were grown. (**e**) Proliferation curve of mouse lymphoid B-cells (Ba/F3) expressing RPL10 wt or p.Arg98Ser. Error bars represent standard deviations of measurements in triplicate. (**f**) Polysome profiling on Ba/F3 cells expressing RPL10 wt or p.Arg98Ser.

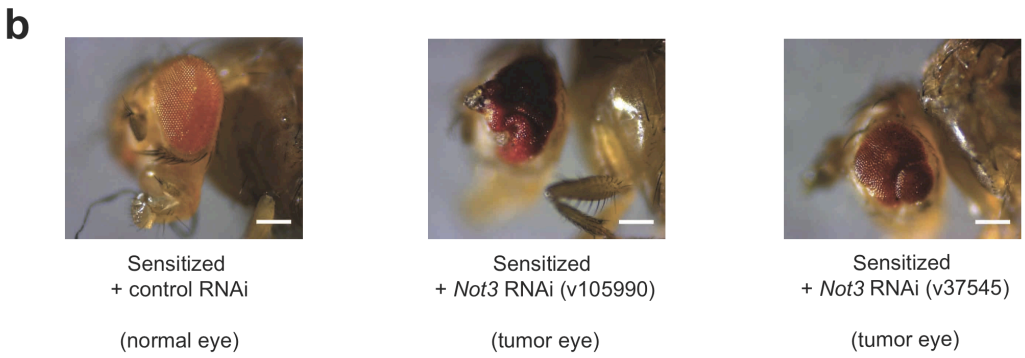


In the context of *CNOT3*, part of the mutations we identified were clearly truncating mutations, while another group of mutations seemed to present missense mutations at residue Arg57. Analysis of mRNA expression, however, revealed that also in the cases with Arg57 mutations, the mutant transcripts are not or weakly expressed **(Supplementary Figure 7)**. This is most likely caused by splicing defects as the mutations are located at the splice donor site of exon 5. The mutations in *CNOT3* thus suggest that this gene acts as a tumor suppressor in T-ALL. CNOT3 is part of the CCR4-NOT complex that regulates gene expression transcriptionally and post-transcriptionally[16]. CNOT3 also mediates self-renewal in mouse embryonic stem cells, where CNOT3 shares many target genes with MYC[17], a known oncogene in T-ALL. To investigate the effect of loss of CNOT3 in tumor formation, we utilized an established *Drosophila melanogaster* eye cancer model. We used the "sensitized" model in which the Notch ligand Delta is overexpressed in the developing eyes. These flies have larger eyes, but by themselves do not develop tumors[3,7,18,19], and this model is relevant for T-ALL given the central role of NOTCH1 signaling in this disease[1]. Reduction of *Not3* expression in this genetic background resulted in a dramatic increase in tumor incidence from 8% of the eyes with the control RNAi to 46% up to 90% with 3 different *Not3* RNAi lines and one line with a P-element insertion in *Not3* (**Fig. 5**). These data support that a reduction of *Not3* expression is sufficient to transform sensitized cells. Using whole exome sequencing, we describe clear differences between pediatric and adult T-ALL and identify a spectrum of driver mutations that function in various cellular processes. One remarkable observation is that a subset of T-ALL cases have accumulated mutations that affect the function of the ribosome, and it is currently unclear what advantage this may provide to the cancer cells. This is, however, very similar to recent findings of deregulated splicing in myelodysplasia and chronic lymphocytic leukemia[20,21,22,23], and may indicate that cancer cells have mechanisms to overcome defects in these basic processes. Indeed, cancer cells may compensate for the deleterious effect of ribosome mutations by acquiring additional mutations, similar to the suppressive effect of the Nmd3 p.Leu291Phe mutation that we describe in the yeast model (**Fig. 4d**). Alternatively, the ribosome mutations may downregulate the hyperactive translation machinery in cancer cells[24], which may be beneficial for the fitness of cancer cells. Our data shed light on the diversity of mutations that are implicated in T-ALL development and on the differences between adult and pediatric T-ALL.

Figure 5. Reduced Not3 expression promotes tumor development in a Drosophila melanogaster sensitized background (next page).

(**a,b**) Sensitized flies overexpressing the *Notch* ligand *Delta* in the eye were crossed to one of three different *Drosophila melanogaster Not3* RNAi fly lines (v105990, v37545, v37547), to the 15271 line with a P-element insertion in *Not3* or to control RNAi flies (RNAi construct against *white* gene). The figure shows quantitative (**a**) and qualitative (**b**) representation of the eye tumor burden in different genotypes. Triple asterisks (***) in **(a)** indicates that tumor incidence in this cross is significantly different from the control cross ($p<0.001$) as analyzed by the 2-tailed Fisher's exact test. Group sizes per cross are indicated in the figure (n). Scale bars in **(b)**: 200 μm.


a
Sensitized + control RNAi (n=266)
92%
8%
Normal eye
Tumor eye (folded)
Sensitized + Not3 RNAi (v105990) (n=246)
10%
90%
(***)
Sensitized + Not3 RNAi (v37545) (n=260)
15%
85%
(***)
Sensitized + Not3 RNAi (v37547) (n=306)
17%
83%
(***)
Sensitized + Not3 P-element (15271) (n=250)
54%
46%
(***)
0
25
50
75
100
% of eyes in cross


b
Sensitized
+ control RNAi
(normal eye)
Sensitized
+ Not3 RNAi (v105990)
(tumor eye)
Sensitized
+ Not3 RNAi (v37545)
(tumor eye)

Accession codes
Sequence and variant data are available via EGA (http://www.ebi.ac.uk/ega/) under accession number EGAS00001000296, and somatic variants are available through a BioMart interface http://lcbmart.aertslab.org/

Acknowledgements
This work was supported by grants from the KU Leuven (concerted action grant to J.C., P.V. and PF/10/016 SymBioSys to J.C., S.A.), the FWO-Vlaanderen (G.0546.11, J.C., P.V., S.A., A.U. and G.0704.11N to S.A.), the Foundation against Cancer (SCIE2006-34, J.C. and 2010-154 to S.A.), an ERC-starting grant (J.C.), the Interuniversity Attraction Poles (IAP) granted by the Federal Office for Scientific, Technical and Cultural Affairs, Belgium (J.C., P.V.), a grant from the Ministry of health, Cancer Plan, (J.C., P.V., S.A.), a grant from the French program Carte d'Identite des Tumeurs (CIT, Ligue Contre le Cancer) and from Canceropole d'Ile de France (J.S.), and from NIH (GM53655 A.J. and S.P.); K.D.K. is a postdoctoral researcher and P.V. is a senior clinical investigator of FWO-Vlaanderen.

Author contributions
All authors contributed to the writing of the manuscript; K.D.K., Z.K.A., N.L., C.B., B.A.H. and A.W.J. designed and performed experiments and analyzed data; C.V. and J.Y. performed and analyzed *Not3* fruit fly experiments; S.P. performed and analyzed *Rpl10* yeast studies; R.L. performed and analyzed polysome profiling experiments; T.G., V.G., E.G., M.P., I.L., G.H., E.C., R.V., B.S., K.J., N.M., I.W., performed experiments and analyzed data; B.C., J.Cloos, J.S., A.U., P.V. collected patient samples and analyzed data; S.A. and J.Cools supervised the project, designed experiments, and analyzed data.

Competing financial interests
The authors declare competing financial interests: author Ning Li is employed by BGI.

METHODS

Cell lines
The Ba/F3 murine pro B cell line and all human T-ALL cell lines were obtained from Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ).

Sequencing
Genomic DNA samples from patients and cell lines were sonicated to fragment size 250 bp on a Bioruptor UCD-300 TM instrument (Diagenode) followed by library preparation using the Truseq DNA sample prep kit (Illumina) and capture with

SeqCapEZ Exome v2.0 (Nimblegen) or Agilent SureSelect 38 Mb (Agilent) capture reagents. Captured samples were sequenced on a Hiseq 2000 (Illumina) operated in paired-end 2 x 100 bp mode. Reads from each sample were aligned to the reference genome version hg19 using BWA[52]. Alignment files were processed further with Genome Analysis ToolKit (GATK) before variant calling. Duplicate removal, local realignment around known INDELs, and base quality recalibration has been done as described in DePristo et al[53]. Variants were called with the Variant Quality Score Recalibration (VQSR) routine of GATK. Variant calling was performed for each sample separately, and somatic variants were identified by selecting the variants observed in diagnosis that were absent in remission. The somatic score for each variant was calculated with the Somatic Sniper[54] algorithm using the processed alignment files for diagnosis and remission samples. INDEL calling was performed with DINDEL[55] using the calibrated bam files for each sample separately. Somatic INDEL detection was performed by identifying the high quality INDELs in the diagnosis sample (minimum read depth of 15 in the variant site, and 15% of the reads supporting the variant allele) and then filtering out the ones observed in the corresponding remission sample (minimum read depth of 1).

Sanger validation of results

All predicted somatic variants shown in Figure 2 were tested and confirmed by conventional Sanger sequencing. These and additional Sanger validated SNVs and INDELs are reported in Supplementary tables 5 and 7. Overall, we performed Sanger validation for 219 SNVs, of which 202 were confirmed (92.2% validation rate) and of 78 INDELs of which 61 were confirmed (78% validation rate). Analysis of Sanger chromatograms was performed using CLC Main Workbench 6 (CLC Bio). Predicted somatic variants that were not confirmed by Sanger sequencing are not shown in Figure 2, but are reported in Supplementary Table 5 and are indicated as false positives. The entire coding sequence of *RPL5*, *RPL10* and *CNOT3* genes was PCR amplified and Sanger sequenced on whole genome amplified DNA from an independent set of 144 T-ALL patients. Primer sequences are provided in Supplementary table 11. Mutation detection was performed using Mutation Surveyor v4.0.4 (Softgenetics) software. Detected variants were confirmed on original, non-amplified material and were tested for their somatic origin on remission DNA if available.

Statistical testing

Throughout this entire work, statistical significance was defined as $p<0.05$. Used statistical testing methods are mentioned in figures and tables.

Drosophila melanogaster experiments

The fly line 15271 (*y^1; P{SUPor-P}l(2)NC136^{KG10496}/CyO*; *ry^{506}*) with a P-element insertion in the *Not3* gene was obtained from the Bloomington Stock Center[56]. *Not3* RNAi Fly lines v37547 (*w1118*;; *P{GD4068}v37547*), v37545 (*w1118*;;

P{GD4068}v37545) and v105990 (;*P{KK102144}VIE-260B*;) were purchased from the Vienna Drosophila RNAi Center (VDRC)[57]. Control RNAi flies (*w*; *Sp/CyO*; *UAS-w dsRNA*) and the Cy8 "sensitized" fly (*w*; *eyeless-Gal4*, *UAS-Dl/ CyO*;) were described previously[3]. All fly strains were crossed to Cy8 virgin females. From the performed crosses, the following F1 progeny was selected as positive:
Cy8 x 15271: ;*eyeless-Gal4*, *UAS-Dl/P{SUPor-P}l(2)NC136*^{KG10496}; *ry*^{506}*/+*
Cy8 x v37547: ;*eyeless-Gal4*, *UAS-Dl/+*; *P{GD4068}v37547/+*
Cy8 x v37545: ;*eyeless-Gal4*, *UAS-Dl/+*; *P{GD4068}v37545/+*
Cy8 x v105990: ;*eyeless-Gal4*, *UAS-Dl*, *P{KK102144}VIE-260B)*;
Cy8 x control: ;*eyeless-Gal4*, *UAS-Dl/CyO*; *UAS-w dsRNA/+*
All flies were raised at 25°C on standard fly food. To analyze the tumor burden, each eye was scored separately in the selected F1 progeny. Eyes were counted as hyperplastic when the eye showed at least one fold.

Yeast experiments
Codon 98 of yeast (*Saccharomyces cerevisiae) RPL10* in the centromeric *LEU2* vector pAJ2522 was changed from AGA to either TCT or TGT, and codon 123 was changed from CAC to CCA by site-specific mutagenesis. Wild type and mutants were introduced into the *Rpl10* deletion strain AJY1437 (*MATα rpl10::KanMX lys2Δ0 met15Δ0 his3Δ1 leu2Δ0 ura3Δ0* pAJ392 - *Rpl10 URA3 CEN*) by plasmid shuffle or the conditional glucose-repressible *Rpl10* strain AJY3373 (*MATα KanMX-GAL1-RPL10 his3Δ1 leu2Δ0 ura3Δ0*) and assayed for growth by plating ten-fold serial dilutions onto selective medium. To test suppression of Rpl10 mutants by mutations in Nmd3, empty vector or vector expressing Nmd3 or Nmd3 p.Leu291Phe was introduced into the indicated strains.

Polysome profiles were analyzed as described[2]. Wild type and *Rpl10* mutants were introduced into AJY1837, containing a glucose repressible *Rpl10* gene (*GAL-Rpl10*), the leptomycin-B (LMB) sensitive allele of *Crm1 p.Thr539Cys* and *Nmd3-GFP,* and AJY2766, containing *GAL-Rpl10* and *Tif6-GFP*. Cultures were grown in selective medium containing galactose. Glucose was added to repress expression of wild type genomic *Rpl10* for 2 hours and LMB was added to a final concentration of 0.1 µg/ml for 30 min to block Nmd3 shuttling. Images were captured using a Nikon E800 microscope fitted with a 100X Plan Apo objective and a Photometrics CoolSNAP ES camera controlled by NIS-Elements AR 2.10 software. Images were prepared using Adobe Photoshop 7.0.

Experiments in mammalian cells
Wild type *RPL10* cDNA (ENST00000344746) was PCR amplified from human thymus cDNA and was cloned into the BglII and EcoRI sites of pMSCV-GFP. The p.Arg98Ser mutation was introduced by mutagenesis of the *RPL10* wild type construct. Mouse lymphoid Pro-B-cells (Ba/F3) were transduced with pMSCV-*RPL10* wild type or p.Arg98Ser retroviral constructs according to standard

methods. To mimic the situation in T-ALL patients which only express mutant and not wild type RPL10 in their leukemia cells (Supplementary Figure 7), Ba/F3 cell experiments were performed with concurrent knock-down of endogenous Rpl10 protein. Delivery of siRNAs into the cells was performed by electoporation on a Gene Pulser Xcell machine (Bio-Rad, Hercules, CA) using exponential decay, 250V, 950 µF. siRNAs against mouse *Rpl10* (MMC.RNAI.N052835.12.2; CCAACAAAUACAUGGUAAAGAGUTG) were used at a concentration of 400 nM during electroporation and were obtained from Integrated DNA Technologies (IDT). Cell proliferation was measured on a Guava flow cytometer (Millipore) at the indicated time points after electroporation.

For polysome profiling, Ba/F3 cells were homogenized in 100mM Tris-HCl pH 7.5, 100mM NaCl, 100mM $MgCl_2$, 1% DOC/Triton-X100, 1mM dithiothreitol, 10ul/ml Protease Inhibitor Cocktail (Sigma), 10ul/ml Phosphatase Inhibitor Cocktail I (Sigma), 10ul/ml Phosphatase Inhibitor Cocktail II (Sigma), 30u/mL RNasin supplemented with 100 ug/ml cycloheximide. After 5 min of incubation on ice, the extract was centrifuged for 5 min at 12,000 g at 4°C. The supernatant was loaded onto a 10-60% (w/v) sucrose gradient and sedimented by centrifugation at 4°C for 150 min at 37,000 rpm in a Beckman SW41 rotor.

REFERENCES

1. Grabher, C., von Boehmer, H. & Look, A.T. Notch 1 activation in the molecular pathogenesis of T-cell acute lymphoblastic leukaemia. *Nat Rev Cancer* **6**, 347-359 (2006).
2. Van Vlierberghe, P. & Ferrando, A. The molecular basis of T cell acute lymphoblastic leukemia. *J. Clin. Invest.* **122**, 3398-3406 (2012).
3. Ferres-Marco, D. et al. Epigenetic silencers and Notch collaborate to promote malignant tumours by Rb silencing. *Nature* **439**, 430-436 (2006).
4. De Keersmaecker, K., Marynen, P. & Cools, J. Genetic insights in the pathogenesis of T-cell acute lymphoblastic leukemia. *Haematologica* **90**, 1116-1127 (2005).
5. Pui, C., Relling, M.V. & Downing, J.R. Acute lymphoblastic leukemia. *N Engl J Med* **350**, 1535-1548 (2004).
6. Homminga, I. et al. Integrated transcript and genome analyses reveal NKX2-1 and MEF2C as potential oncogenes in T cell acute lymphoblastic leukemia. *Cancer Cell* **19**, 484-497 (2011).
7. Ntziachristos, P. et al. Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. *Nat. Med.* **18**, 298-301 (2012).
8. Zhang, J. et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481**, 157-163 (2012).
9. Larson, D.E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311-317 (2012).
10. Dees, N.D. et al. MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589-1598 (2012).
11. Narla, A. & Ebert, B.L. Ribosomopathies: human disorders of ribosome dysfunction. *Blood* **115**, 3196-3205 (2010).
12. Rao, S. et al. Inactivation of ribosomal protein L22 promotes transformation by induction of the stemness factor, Lin28B. *Blood* **120**, 3764-3773 (2012).
13. Hofer, A., Bussiere, C. & Johnson, A.W. Mutational analysis of the ribosomal protein Rpl10 from yeast. *J. Biol. Chem.* **282**, 32630-32639 (2007).
14. Hedges, J. et al. Release of the export adapter, Nmd3p, from the 60S ribosomal subunit requires Rpl10p and the cytoplasmic GTPase Lsg1p. *EMBO J* **24**, 567-579 (2005).
15. Lo, K. et al. Defining the pathway of cytoplasmic maturation of the 60S ribosomal subunit. *Mol. Cell* **39**, 196-208 (2010).
16. Collart, M.A. & Panasenko, O.O. The Ccr4--not complex. *Gene* **492**, 42-53 (2012).
17. Hu, G. et al. A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes. Dev.* **23**, 837-848 (2009).
18. Palomero, T. et al. Mutational loss of PTEN induces resistance to NOTCH1 inhibition in T-cell leukemia. *Nat. Med.* **13**, 1203-1210 (2007).
19. Bossuyt, W. et al. The atonal proneural transcription factor links

differentiation and tumor formation in Drosophila. *PLoS Biol.* **7**, e40 (2009).
20. Quesada, V. et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet* **44**, 47-52 (2012).
21. Graubert, T.A. et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat Genet* **44**, 53-57 (2012).
22. Yoshida, K. et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64-69 (2011).
23. Papaemmanuil, E. et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med* **365**, 1384-1395 (2011).
24. Ruggero, D. & Pandolfi, P.P. Does the ribosome translate cancer? *Nat Rev Cancer* **3**, 179-192 (2003).
25. Weng, A.P. et al. Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* **306**, 269-271 (2004).
26. Fabbri, G. et al. Analysis of the chronic lymphocytic leukemia coding genome: role of NOTCH1 mutational activation. *J Exp Med* **208**, 1389-1401 (2011).
27. Puente, X.S. et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**, 101-105 (2011).
28. Westhoff, B. et al. Alterations of the Notch pathway in lung cancer. *Proc Natl Acad Sci USA* **106**, 22293-22298 (2009).
29. Agrawal, N. et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* **333**, 1154-1157 (2011).
30. Stransky, N. et al. The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-1160 (2011).
31. Robinson, D.R. et al. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nat. Med.* **17**, 1646-1651 (2011).
32. Jiao, X. et al. Somatic mutations in the Notch, NF-KB, PIK3CA, and Hedgehog pathways in human breast cancers. *Genes Chromosomes Cancer* **51**, 480-489 (2012).
33. Thompson, B.J. et al. The SCFFBW7 ubiquitin ligase complex as a tumor suppressor in T cell leukemia. *J Exp Med* **204**, 1825-1835 (2007).
34. Welcker, M. & Clurman, B.E. FBW7 ubiquitin ligase: a tumour suppressor at the crossroads of cell division, growth and differentiation. *Nat Rev Cancer* **8**, 83-93 (2008).
35. Tosello, V. et al. WT1 mutations in T-ALL. *Blood* **114**, 1038-1045 (2009).
36. King-Underwood, L. & Pritchard-Jones, K. Wilms' tumor (WT1) gene mutations occur mainly in acute myeloid leukemia and may confer drug resistance. *Blood* **91**, 2961-2968 (1998).
37. De Keersmaecker, K. et al. The TLX1 oncogene drives aneuploidy in T cell transformation. *Nat. Med.* **16**, 1321-1327 (2010).
38. Klauck, S.M. et al. Mutations in the ribosomal protein gene RPL10 suggest a

novel modulating disease mechanism for autism. *Mol. Psychiatry* **11**, 1073-1084 (2006).

39. Chiocchetti, A. et al. Mutation and expression analyses of the ribosomal protein gene RPL10 in an extended German sample of patients with autism spectrum disorder. *Am. J. Med. Genet. A* **155A**, 1472-1475 (2011).
40. Kalender Atak, Z. et al. High accuracy mutation detection in leukemia on a selected panel of cancer genes. *PLoS ONE* **7**, e38463 (2012).
41. Durieux, A., Prudhon, B., Guicheney, P. & Bitoun, M. Dynamin 2 and human diseases. *J. Mol. Med.* **88**, 339-350 (2010).
42. Van Vlierberghe, P. et al. PHF6 mutations in T-cell acute lymphoblastic leukemia. *Nat Genet* **42**, 338-342 (2010).
43. Van Vlierberghe, P. et al. ETV6 mutations in early immature human T cell leukemias. *J Exp Med* **208**, 2571-2579 (2011).
44. Lower, K.M. et al. 1024C> T (R342X) is a recurrent PHF6 mutation also found in the original Börjeson-Forssman-Lehmann syndrome family. *Eur. J. Hum. Genet.* **12**, 787-789 (2004).
45. Ono, R. et al. LCX, leukemia-associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with trilineage dysplasia having t(10;11)(q22;q23). *Cancer Research* **62**, 4075-4080 (2002).
46. Lorsbach, R.B. et al. TET1, a member of a novel protein family, is fused to MLL in acute myeloid leukemia containing the t(10;11)(q22;q23). *Leukemia* **17**, 637-641 (2003).
47. Burmeister, T. et al. The MLL recombinome of adult CD10-negative B-cell precursor acute lymphoblastic leukemia: results from the GMALL study group. *Blood* **113**, 4011-4015 (2009).
48. van Haaften, G. et al. Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nat Genet* **41**, 521-523 (2009).
49. Gui, Y. et al. Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nat Genet* **43**, 875-878 (2011).
50. Lederer, D. et al. Deletion of KDM6A, a histone demethylase interacting with MLL2, in three patients with Kabuki syndrome. *Am J Hum Genet* **90**, 119-124 (2012).
51. Pasqualucci, L. et al. Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat Genet* **43**, 830-837 (2011).
52. Li, H., Li, H., Durbin, R. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
53. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491-498 (2011).
54. Larson, D.E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311-317 (2012).
55. Albers, C.A. et al. Dindel: Accurate indel calls from short-read data. *Genome Res.* **21**, 961-973 (2011).

56. Bellen, H.J. et al. The BDGP gene disruption project: single transposon insertions associated with 40% of Drosophila genes. *Genetics* **167**, 761-781 (2004).
57. Dietzl, G. et al. A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. *Nature* **448**, 151-156 (2007).

SUPPLEMENTARY INFORMATION FOR:

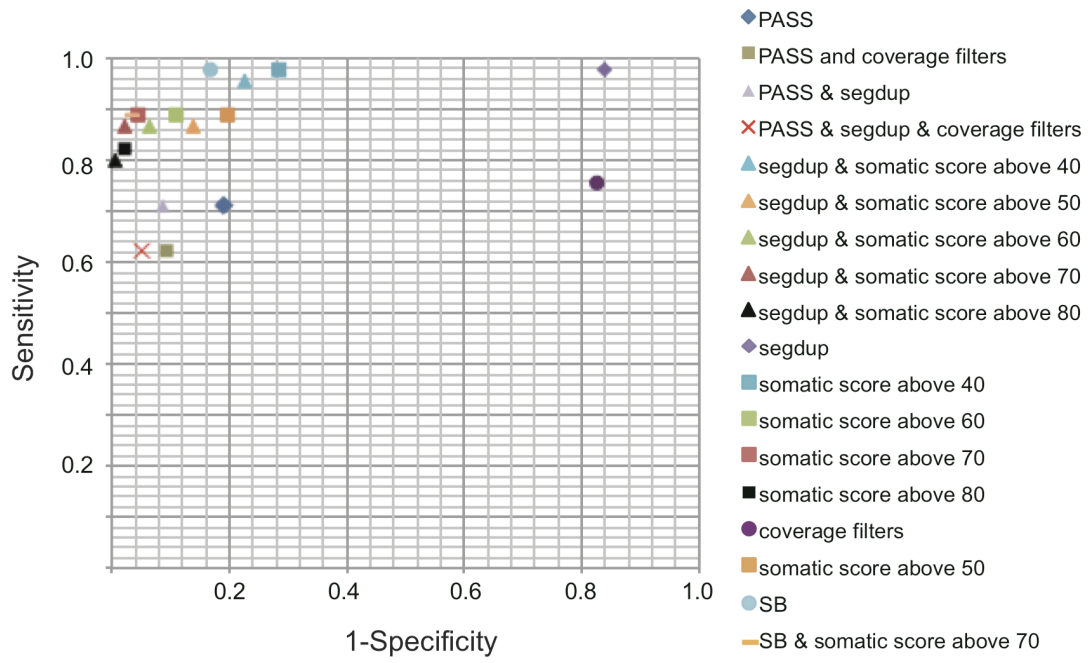
"Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia"

SUPPLEMENTARY NOTE

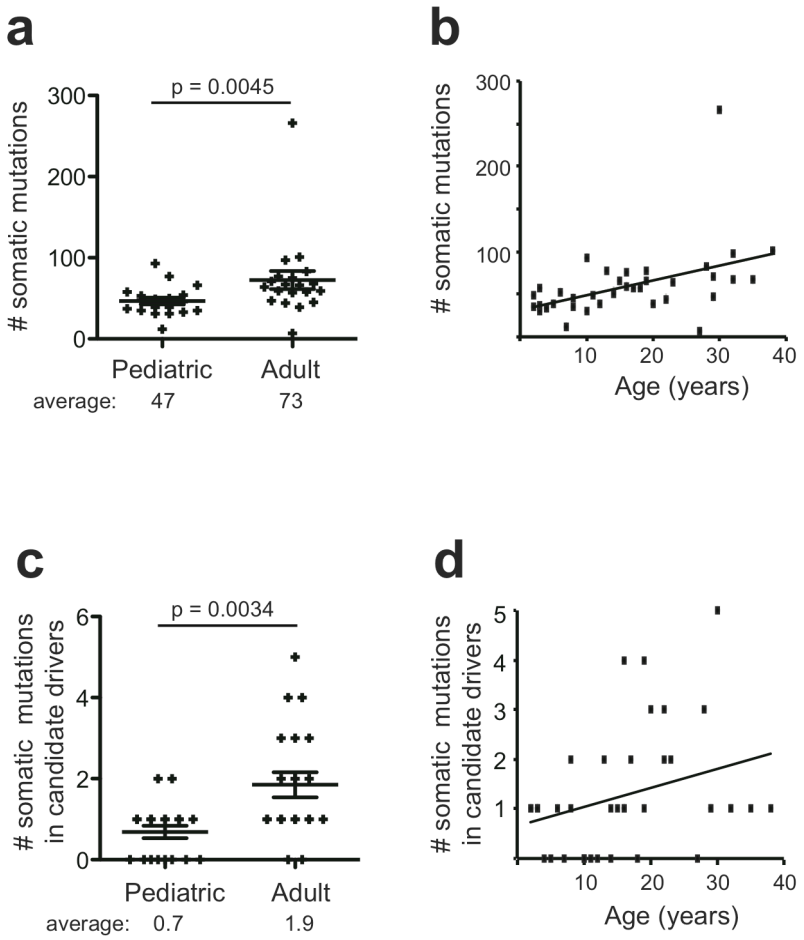
Leukemia samples

T-ALL patient samples were collected at various institutions. All patients have given their informed consent and all samples were obtained according to the guidelines of the local ethical committees. This study was approved by the ethical committee of the University Hospital Leuven. Diagnosis of T-ALL was based on morphology, cytochemistry and immunophenotyping according to the World Health Organization and European Group for the Immunological Characterization of Leukemias criteria. A detailed description of the clinico-biological characteristics of the analyzed patient samples is provided in Supplementary table 1.

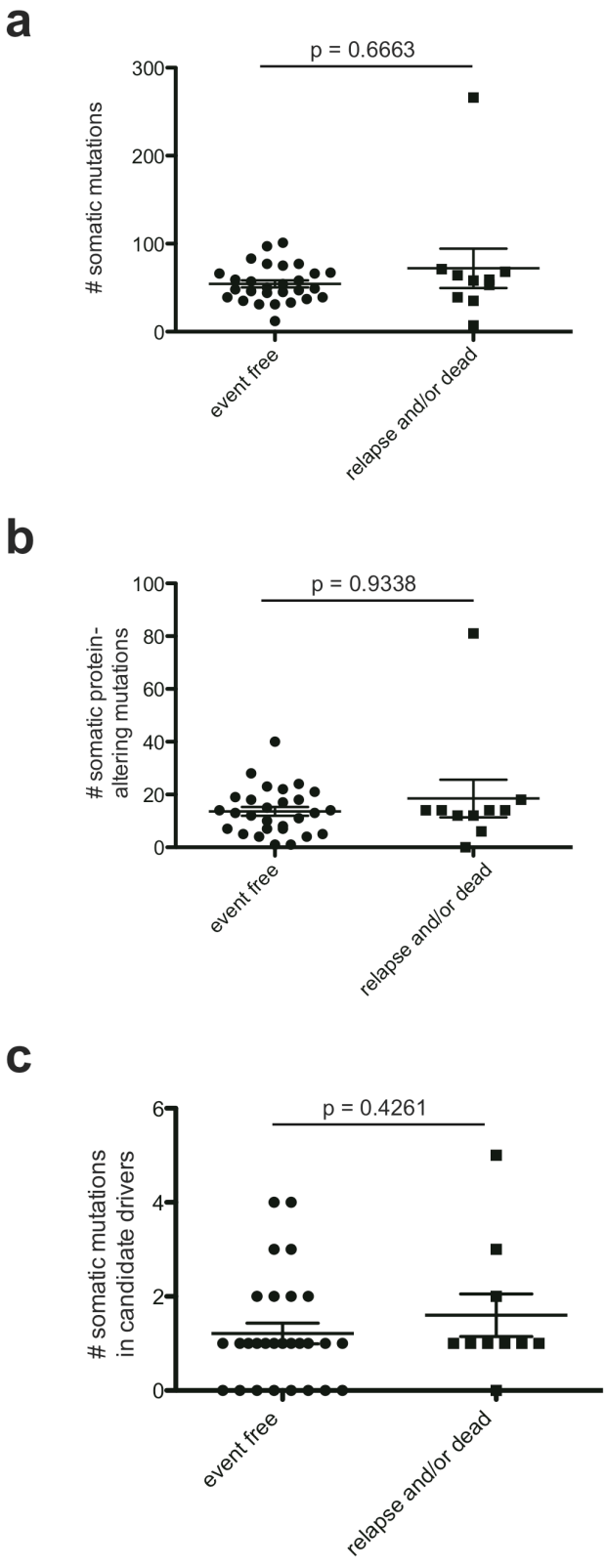
Supplemental Figure 1. Comparison of somatic SNV filtering strategies.
Variant quality filtering is denoted as 'PASS' (referring to the quality tag for high quality variants predicted by VQSR). Coverage filtering is defined as depth of coverage ≥ 10 reads in diagnosis and remission, and variant allele frequency ≥ 20% in diagnosis and ≤ 5% in remission. For repeat region filtering ('segdup' in the legend), all variants in segmental duplication regions are removed. The final filtering strategy used in the rest of this work is indicated with a red square.



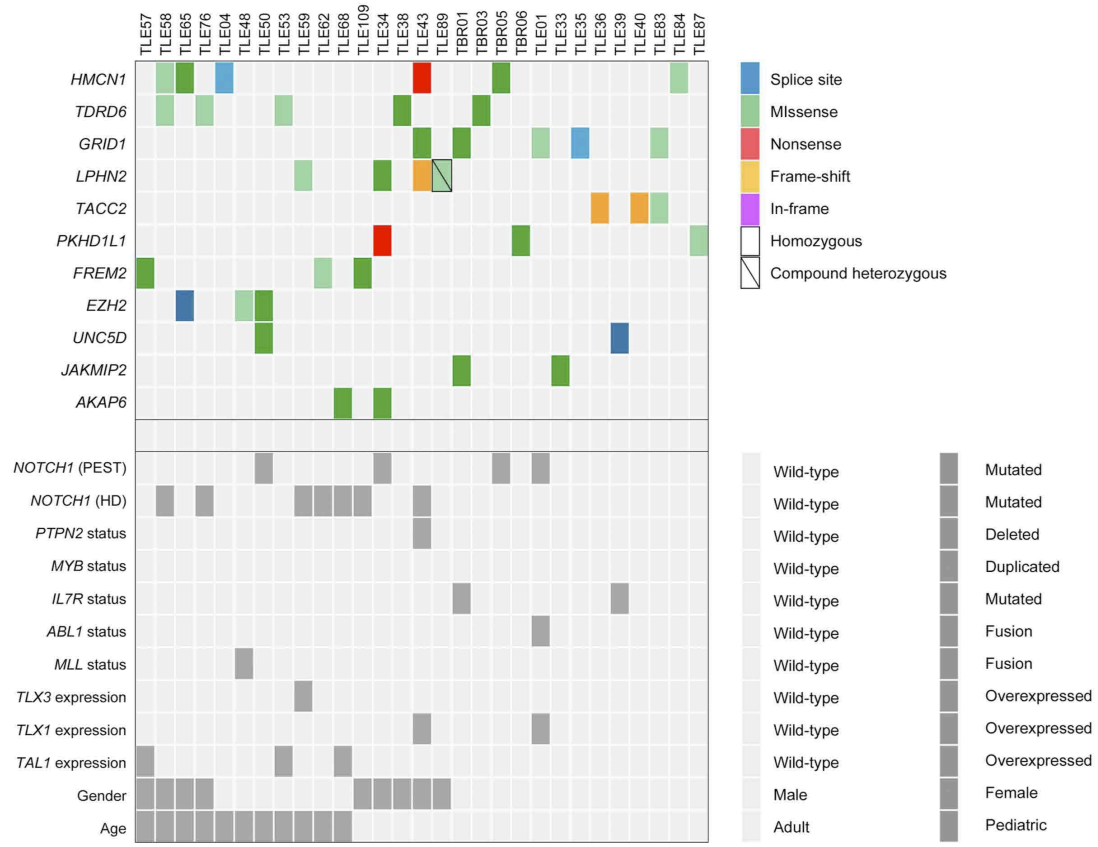
Supplemental Figure 2. The number of somatic mutations increases with age.
(**a**) Box plot showing the number of somatic mutations in pediatric (age ≤ 15 years) and in adult (age ≥ 16) T-ALL samples. Average and s.e.m. is indicated on the plots. The reported p-value tests whether there is a significant difference between mutation number in adults versus children and was calculated using a 2-tailed Wilcoxon signed rank test. Group size pediatric: n=19; adult: n=20. (**b**) Dot plot representing the number of somatic mutations versus patient age. (**c**) Box plot showing the number of somatic mutations in candidate driver genes in pediatric and in adult T-ALL patients. Average and s.e.m. is indicated on the plots. The reported p-value tests whether there is a significant difference between mutation number in adults versus children and was calculated using a 2-tailed Wilcoxon signed rank test. Group size pediatric: n=19; adult: n=20. (**d**) Dot plot representing the number of somatic mutations in candidate driver genes versus patient age.



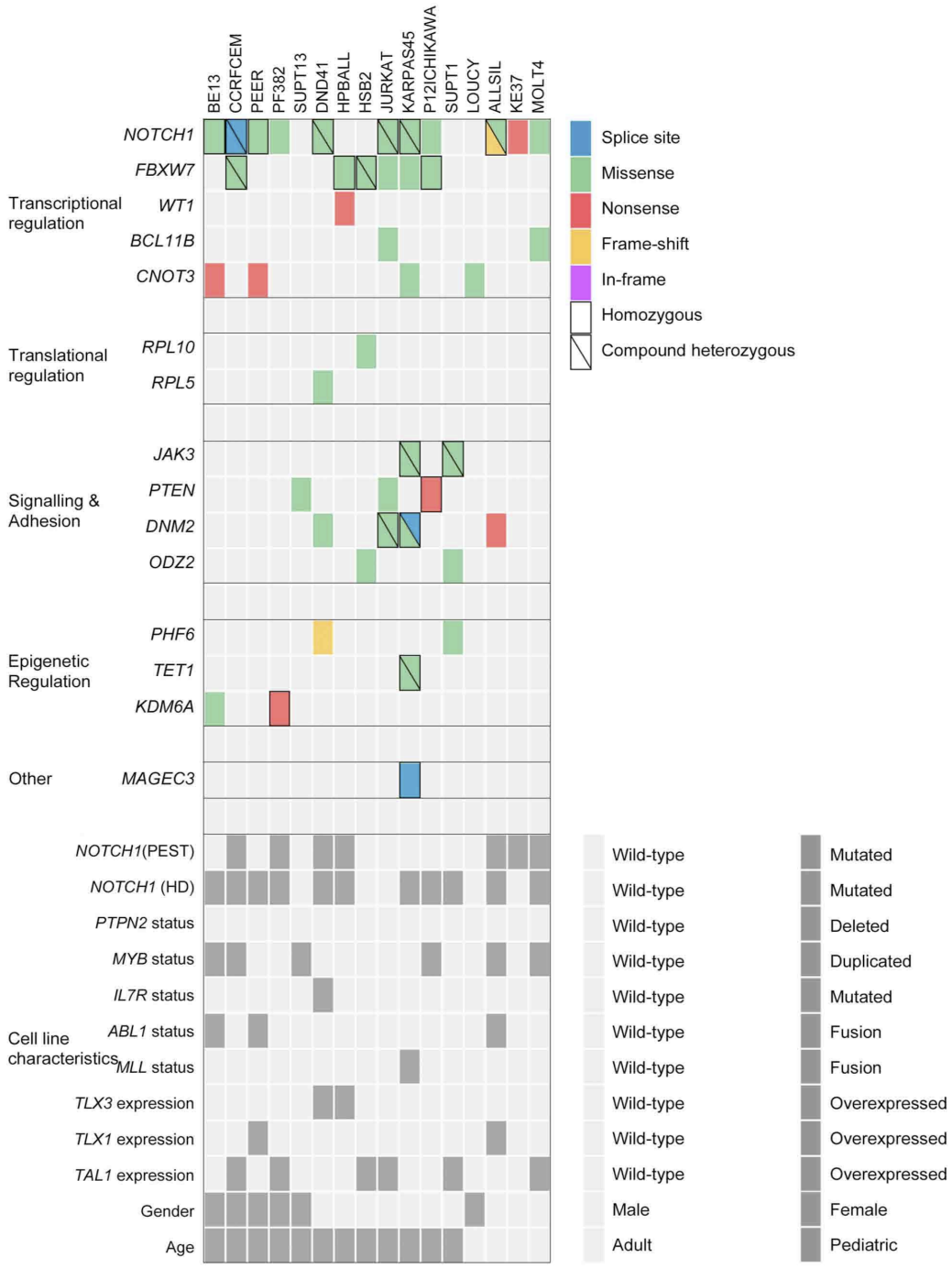
Supplemental Figure 3. The number of mutations does not correlate with outcome.
(**a-c**) Box plots showing number of somatic mutations (**a**), number of protein-altering somatic mutations (**b**) and number of somatic mutations in candidate driver genes (**c**) in patients that did not undergo an event versus patient that relapsed and/or died due to T-ALL. Average and s.e.m. is indicated on the plots. All reported p-values test whether there is a significant difference between mutation number in event free versus relapsed and/or leukemia induced dead patiens and were calculated using a 2-tailed Wilcoxon singed rank test. Group size event free: n=28; relapse and/or dead: n=10.



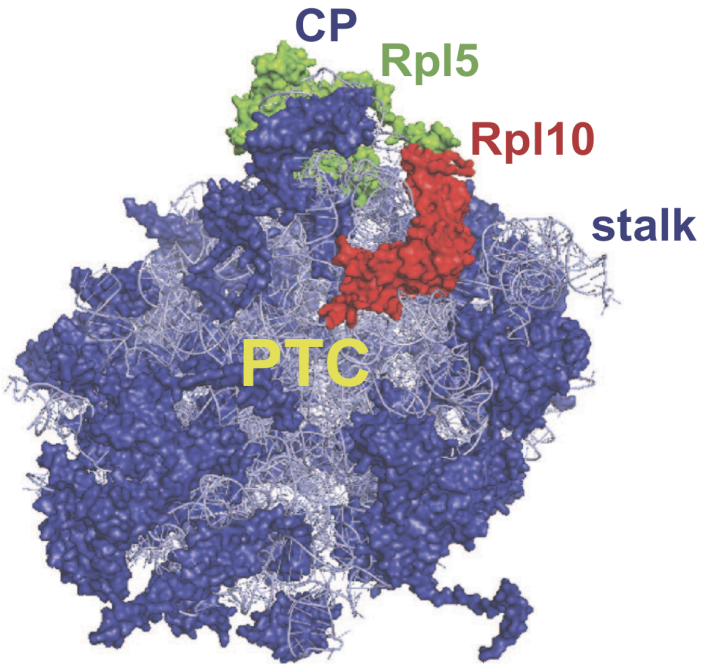
Supplemental Figure 4. Overview of mutations in recurrent but not significantly mutated genes in T-ALL samples.
Mutations observed in 11 genes that were recurrently mutated in at least 2 of the diagnosis-remission pairs but that were not significantly mutated as determined by Genome MuSiC. For clarity, only samples harboring mutations in any of these 11 genes are shown in the figure. Each type of mutation is indicated with a different color as indicated in the legend and symbols for homozygous and compound heterozygous mutations are explained. Mutations with no indication are heterozygous. All mutations shown in this figure were validated by conventional Sanger sequencing. Relevant patient characteristics (identified by Sanger sequencing, karyotyping, or gene expression) are included at the bottom of the heatmap. Mutations in *NOTCH1* were hard to identify by exome sequencing due to the low coverage of *NOTCH1* exons. *NOTCH1* mutations detected by Sanger sequencing are indicated at the bottom of the heatmap under patient characteristics.



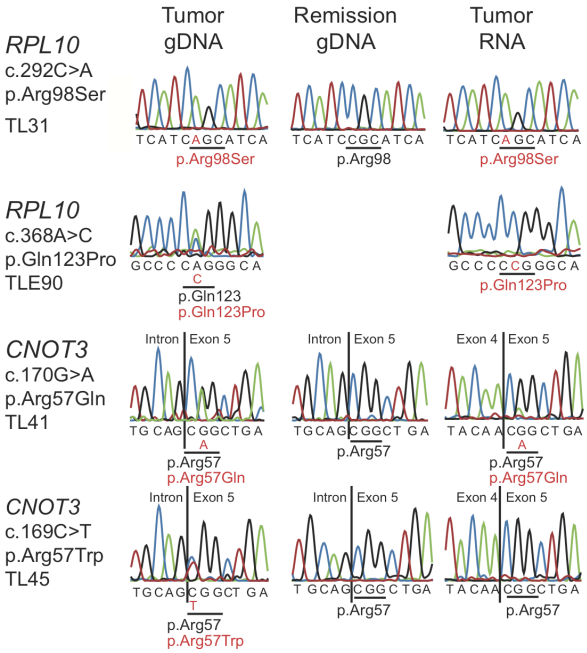
Supplemental Figure 5. Overview of mutations in 15 identified candidate T-ALL driver.
Mutations observed in the 15 selected candidate T-ALL driver genes in 17 sequenced T-ALL cell lines. Each type of mutation is indicated with a different color as indicated in the legend and symbols for homozygous and compound heterozygous mutations are explained. Mutations with no indication are heterozygous. All mutations shown in this figure were validated by conventional Sanger sequencing. Relevant cell line characteristics (identified by Sanger sequencing, karyotyping, or gene expression) are included at the bottom of the heatmap. Mutations in *NOTCH1* were hard to identify by exome sequencing due to the low coverage of *NOTCH1* exons. *NOTCH1* mutations detected by Sanger sequencing are indicated at the bottom of the heatmap under cell line characteristics.



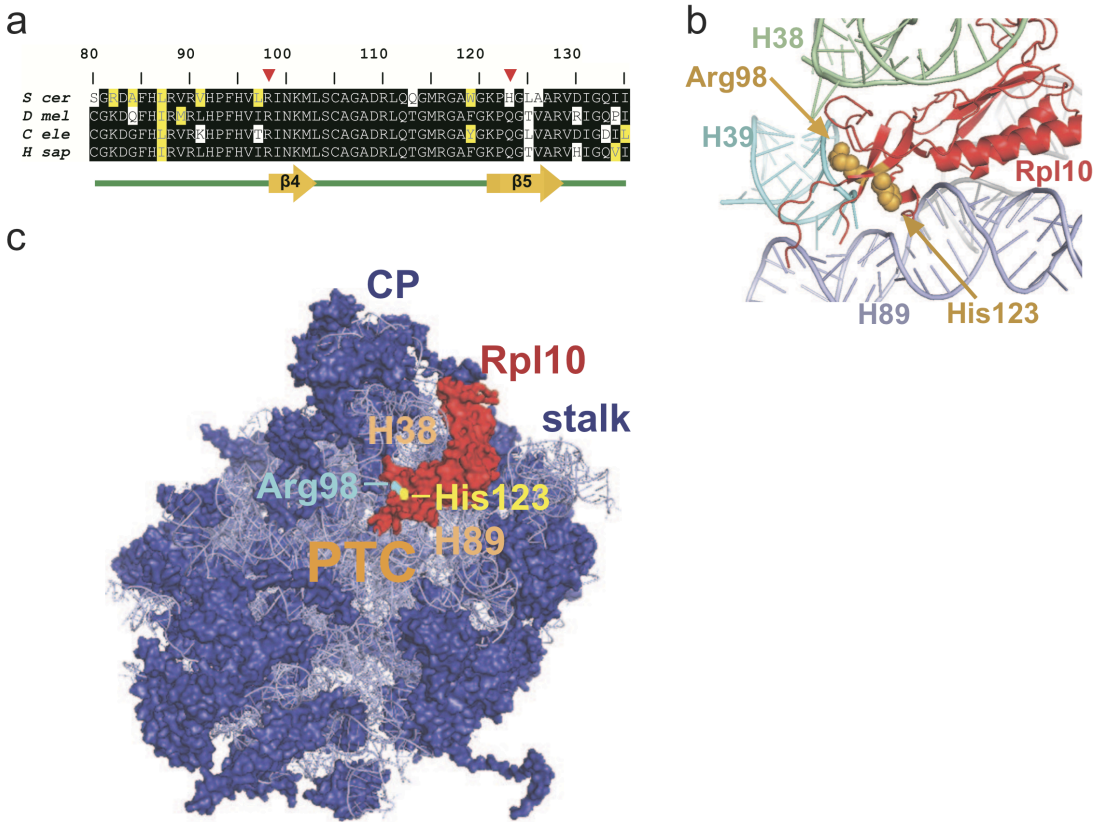
Supplemental Figure 6. Localization of Rpl10 and Rpl5 proteins in the 60S ribosomal subunit. The figure shows the 'crown view' of the rRNA and proteins of the yeast 60S ribosomal subunit with indication of Rpl10 and Rpl5 proteins. PTC: peptidyltransferase center; CP: central protuberance. This model is based on the yeast 3Å crystal structure (PDB entries 3U5E and 3U5D).



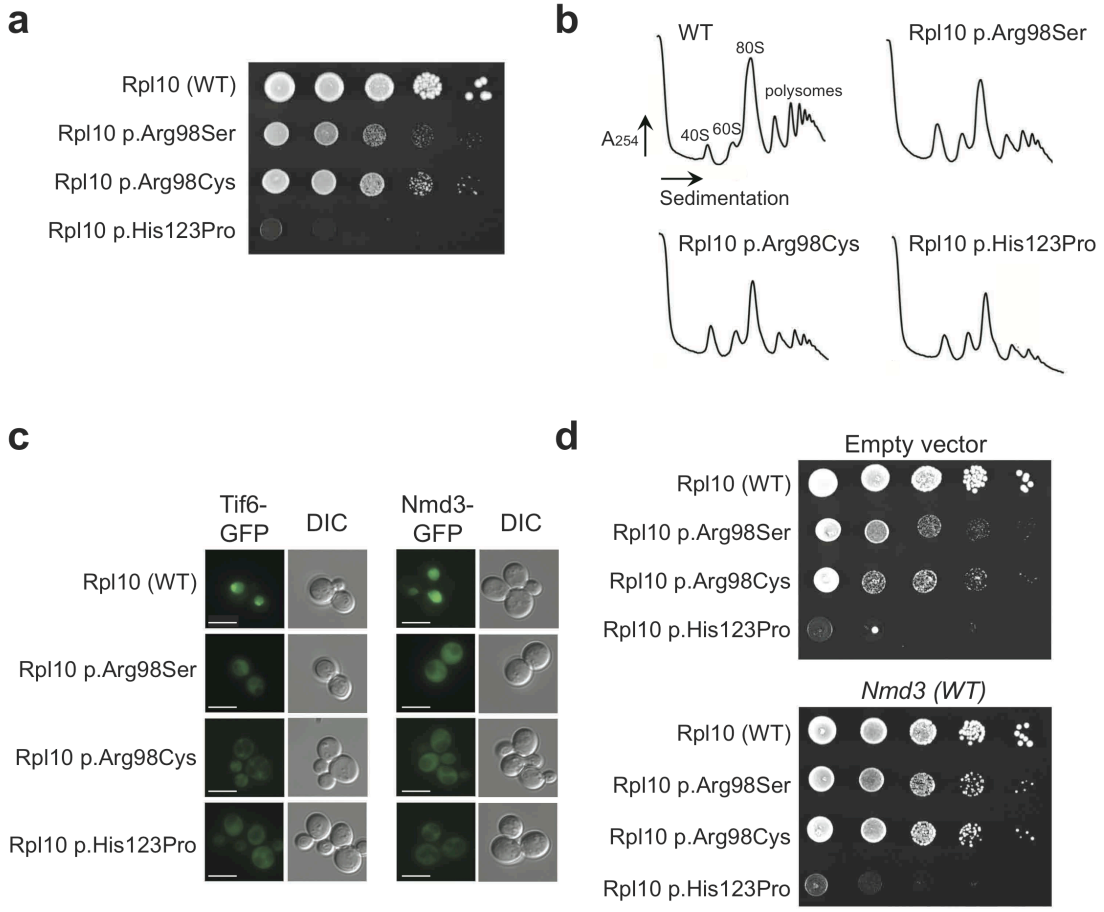
Supplemental Figure 7. DNA and RNA Sanger results RPL10 and CNOT3 mutations. Representative chromatograms illustrating presence or absence of indicated *RPL10* or *CNOT3* mutations in diagnostic or remission genomic DNA or diagnostic RNA.



Supplemental Figure 8. RPL10 mutated residues are close to the catalytic center of the ribosome. (**a**) Alignment of Rpl10 protein sequence from yeast (*S cer*), Drosophila (*D mel*), nematodes (*C ele*) and human (*H sap*). Amino acid positions are given and Arg98 and Gln123 (or His123 in yeast) are indicated (▼). β-sheets in the secondary structure of Rpl10 are indicated under the alignment. (**b**) 3D representation of the region of yeast Rpl10 containing Arg98 and His123 and surrounding rRNA helices H38, H38 and H89. The figure shows that Rpl10 residues Arg98 and His123 are opposing residues in a beta-hairpin loop of Rpl10. (**c**) Model showing the position of Rpl10 and its residues Arg98 and His123 into the 60S yeast ribosomal subunit ('crown view'). The figure illustrates that Rpl10, and particularly residues Arg98 and His123, are close to the peptidyltransferase center (PTC) in the ribosome. For Reference, rRNA helices H38 and H89 that are also shown in panel (b) are indicated. CP: Central Protuberance. Modeling in panels (b) and (c) is based on the yeast 3Å crystal structure (PDB entries 3U5E and 3U5D).



Supplemental Figure 9. T-ALL associated RPL10 mutants impair proliferation and ribosome biogenesis in yeast cells.
(**a**) Growth of conditional glucose-repressible Rpl10 yeast cells expressing wild-type (WT) Rpl10, or the T-ALL associated Rpl10 p.Arg98Ser, Rpl10 p.Arg98Cys and Rpl10 p.His123Pro mutants was compared by plating ten-fold serial dilutions. (**b**) Polysome profiles for Rpl10 WT, Rpl10 p.Arg98Ser, Rpl10 p.Arg98Cys and Rpl10 p.His123Pro yeast cells were carried out as described for Fig 4b with the exception that the Rpl10 alleles were expressed in a glucose-repressible Rpl10 strain shifted to glucose for six hours. (**c**) Fluorescence microscopy of Tif6-GFP or Nmd3-GFP for the indicated yeast cells was done as described in the legend to Fig 4c. Scale bars: 5 μm. (**d**) Rpl10 WT, Rpl10 p.Arg98Ser, Rpl10 p.Arg98Cys and Rpl10 p.His123Pro yeast cells were transformed with empty vector or vector WT Nmd3. Ten-fold serial dilutions were grown.



Supplementary Table 1. Description patient set exome sequencing.
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Supplementary Table 2. Average alignment results over 123 samples

	Average ± St. Dev.	Range
Number of high-quality aligned bases (Gbp)	7.1 ± 2.5	3.9 – 17.2
Mean target coverage	109.12 ± 31.8	51.3 - 239
Percentage of bases covered by at least 2 reads (%)	95.7 ± 0.8	92.5 – 97.3
Percentage of bases covered by at least 10 reads (%)	91.0 ± 2.0	81.5 – 95.7
Percentage of bases covered by at least 20 reads (%)	84.9 ± 3.8	68.7 – 94.1

Supplementary Table 3. Sequencing and mapping statistics of the 123 samples.
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Supplementary Table 4. Detected SNV and INDEL numbers.

Supplementary Table 4A. Number of somatic SNVs and INDELs observed in 39 samples

Sample	# somatic SNVs	# somatic INDELs	Total # of somatic variations	#of protein altering somatic SNVs	#of protein altering somatic INDELs	Total # of protein altering somatic variations	
TLE02	28	7	35	6	0	6	
TLE03	35	4	39	13	1	14	
TLE07	4	27	31	0	5	5	
TLE09	5	7	12	0	1	1	
TLE10	51	17	68	14	4	18	
TLE29	24	9	33	1	0	1	
TLE31	37	2	39	4	0	4	
TLE32	9	84	93	0	9	9	
TLE33	40	7	47	17	1	18	
TLE34	66	153	66	19	61	19	*
TLE36	27	32	59	6	4	10	
TLE38	33	12	45	17	5	22	
TLE39	53	5	58	20	1	21	
TLE40	38	63	101	0	14	14	
TLE41	53	6	59	11	1	12	
TLE42	64	253	64	14	27	14	*
TLE43	250	16	266	76	5	81	
TLE44	46	25	71	12	2	14	
TLE45	37	7	44	10	2	12	
TLE50	30	9	39	13	0	13	
TLE51	40	13	53	10	4	14	
TLE54	61	5	66	13	1	14	
TLE55	35	2	37	8	0	8	
TLE57	35	316	35	13	103	13	*
TLE60	30	28	58	6	6	12	

TLE61	74	3	77	15	0	15	
TLE63	44	5	49	4	3	7	
TLE64	44	2	46	11	0	11	
TLE65	35	19	54	6	1	7	
TLE66	34	14	48	2	3	5	
TLE67	47	3	50	7	0	7	
TLE68	23	8	31	3	1	4	
TLE109	69	14	83	20	4	24	
TLE110	0	7	7	0	0	0	
TBR01	63	4	67	26	2	28	
TBR03	52	5	57	17	0	17	
TBR05	86	11	97	37	3	40	
TBR06	45	30	75	11	7	18	
TBR08	63	14	77	20	3	23	

* the INDEL calls from these samples are excluded from further analysis

Supplementary Table 4B. Number of SNVs and INDELs observed in 28 diagnosis-only samples

Sample	# of SNVs	# of INDELs	Total # of variation	# of protein altering SNVs	# of protein altering INDELs	Total # of protein altering variations
TLE01	3683	58	3741	274	7	281
TLE04	3433	44	3477	179	6	185
TLE05	3959	45	4004	193	8	201
TLE06	3602	66	3668	201	12	213
TLE08	3791	42	3833	198	4	202
TLE35	12118	132	12250	280	11	291
TLE37	5704	32	5736	256	6	262
TLE47	5983	40	6023	263	7	270
TLE48	9229	50	9279	298	5	303
TLE53	6040	40	6080	271	2	273
TLE56	6306	49	6355	303	6	309
TLE58	14013	114	14127	401	18	419
TLE59	8323	33	8356	333	8	341
TLE62	5326	47	5373	232	6	238
TLE76	16973	64	17037	262	5	267
TLE78	14446	124	14570	251	12	263
TLE79	9487	108	9595	194	16	210
TLE80	5052	80	5132	211	8	219
TLE82	5922	86	6008	221	8	229
TLE83	6097	117	6214	212	15	227
TLE84	8393	123	8516	236	18	254
TLE85	8717	113	8830	219	11	230
TLE87	4769	77	4846	188	8	196
TLE89	8113	110	8223	236	15	251
TLE90	25753	494	26247	996	117	1113
TLE91	8047	102	8149	217	10	227
TLE92	5400	90	5490	212	10	222
TBR09	5489	100	5589	307	8	315

Supplementary Table 4C. Number of SNVs and INDELs observed in 17 T-ALL cell lines

Sample	# of SNVs	# of INDELs	Total # of variation	# of protein altering SNVs	# of protein altering INDELs	Total # of protein altering variations
BE13	7127	58	7185	553	17	570
HSB2	10400	281	10681	1522	162	1684
KARPAS45	23897	110	24007	6094	43	6137
SUPT1	18056	371	18427	3481	211	3692
ALLSIL	6099	52	6151	424	11	435
DND41	13220	230	13450	2295	122	2417
HPBALL	4772	49	4821	519	13	532
JURKAT	21269	517	21786	4627	309	4936
KE37	5863	50	5913	296	8	304
MOLT4	9195	123	9318	1190	47	1237
P12ICHIKAWA	5475	54	5529	405	11	416
PEER	6342	38	6380	420	7	427
SUPT13	4932	50	4982	268	11	279
TALL1	6343	38	6381	347	12	359
PF382	9640	63	9703	1705	18	1723
CCRFCEM	12838	544	13382	2735	234	2969
LOUCY	4185	63	4248	292	9	301

Supplementary Table 5. Protein-altering (A) somatic SNVs and (B) somatic INDELs in 39 diagnosis-remission pairs.

The table is not included in the thesis due to space constrains and can be obtained through the published article.

Supplementary Table 6. 20 significantly mutated genes.

Gene	SNVs	INDELs	p-value FCPT	p-value LRT	p-value CT	FDR FCPT	FDR LRT	FDR CT
PHF6	6	3	0	0	0	0	0	0
FBXW7	5	0	3E-07	4E-12	6E-12	2E-03	4E-08	6E-08
WT1	1	2	7E-04	9E-08	1E-07	1E+00	6E-04	4E-04
JAK3	3	1	2E-04	2E-07	7E-08	1E+00	1E-03	4E-04
DNM2	3	1	3E-04	3E-07	1E-07	1E+00	1E-03	4E-04
PTEN	2	1	7E-04	8E-07	1E-07	1E+00	2E-03	4E-04
CNOT3	3	1	6E-04	9E-07	2E-07	1E+00	2E-03	6E-04
NOTCH1	4	0	4E-04	3E-06	2E-07	1E+00	5E-03	6E-04
BCL11B	3	0	2E-03	4E-07	6E-07	1E+00	1E-03	1E-03
TET1	2	2	1E-03	2E-06	6E-07	1E+00	3E-03	1E-03
KDM6A	1	2	3E-03	6E-07	1E-06	1E+00	2E-03	2E-03
PPP1R15A *	0	2	3E-02	2E-06	2E-05	1E+00	4E-03	3E-02
ODZ2	3	0	2E-02	1E-05	2E-05	1E+00	1E-02	3E-02
TLR1	1	1	4E-02	3E-05	3E-05	1E+00	4E-02	4E-02
RPL10	2	0	4E-02	3E-06	4E-05	1E+00	5E-03	5E-02
ST8SIA2 *	2	0	7E-02	8E-06	1E-04	1E+00	1E-02	1E-01
RPL5	2	0	7E-02	1E-04	1E-04	1E+00	9E-02	1E-01
FAM55A *	1	0	3E-01	1E-04	2E-04	1E+00	9E-02	2E-01
MAGEC3	2	0	1E-01	2E-05	2E-04	1E+00	2E-02	2E-01
MTMR8 *	2	0	1E-01	2E-04	2E-04	1E+00	1E-01	2E-01

* the gene is not recurrently mutated across different samples (the gene has either one mutation, or several mutations in the same sample).

Supplementary Table 7. Protein-altering SNVs and INDELs in 15 candidate genes in (A) diagnosis-only samples and (B) cell lines.
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Supplementary Table 8. Patient characteristics *CNOT3* (NM_014516.3) mutated cases.
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Supplementary Table 9. Patient characteristics RPL mutated cases.
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Supplementary Table 10. Association between candidate driver mutations and clinicobiological features

AGE		# adult (≥16 y) patients	# pediatric (≤15) patients	p-value 2-tailed Fisher's test
NOTCH1	mutant	15	14	p=0.6300
n=67	wild type	22	16	
FBXW7	mutant	7	1	p=0.0657
n=67	wild type	30	29	
WT1	mutant	2	5	p=0.4334
n=67	wild type	35	25	
BCL11B	mutant	4	1	p=0.3700
n=67	wild type	33	29	
CNOT3	mutant	7	1	**p=0.0107**
n=211	wild type	82	121	
RPL10	mutant	1	10	**p=0.0268**
n=211	wild type	88	112	
RPL5	mutant	2	2	p=1.0000
n=211	wild type	87	120	
JAK3	mutant	4	3	p=1.0000
n=67	wild type	33	27	
PTEN	mutant	3	1	p=0.6220
n=67	wild type	34	29	
DNM2	mutant	1	3	p=0.3179
n=67	wild type	36	27	
ODZ2	mutant	1	1	p=1.0000
n=67	wild type	36	29	
PHF6	mutant	11	1	**p=0.0083**
n=67	wild type	26	29	
TET1	mutant	3	1	p=0.6220
n=67	wild type	34	29	
KDM6A	mutant	3	0	p=0.2469
n=67	wild type	34	30	
MAGEC3	mutant	2	0	p=0.4980
n=67	wild type	35	30	

NOTCH1 status		# NOTCH1 wt patients	# NOTCH1 mut patients	p-value 2-tailed Fisher's test
FBXW7	mutant	5	3	p=1.0000
n=67	wild type	33	26	
WT1	mutant	4	3	p=1.0000
n=67	wild type	34	26	
BCL11B	mutant	0	5	**p=0.0123**
n=67	wild type	38	24	
CNOT3	mutant	2	6	p=0.1369
n=84	wild type	43	33	
RPL10	mutant	3	6	P=0.2918

n=84	wild type	42	33	
RPL5	mutant	1	2	p=0.5948
n=84	wild type	44	37	
JAK3	mutant	5	2	p=0.6899
n=67	wild type	33	27	
PTEN	mutant	4	0	p=0.1273
n=67	wild type	34	29	
DNM2	mutant	2	2	p=1.0000
n=67	wild type	36	27	
ODZ2	mutant	0	2	p=0.1836
n=67	wild type	38	27	
PHF6	mutant	4	8	p=0.1078
n=67	wild type	34	21	
TET1	mutant	2	2	p=1.0000
n=67	wild type	36	27	
KDM6A	mutant	1	2	p=0.5744
n=67	wild type	37	27	
MAGEC3	mutant	2	0	p=0.5016
n=67	wild type	36	29	

Supplementary Table 11. *RPL10*, *RPL5* and *CNOT3* PCR and sequencing primers
The table is not included in the thesis due to space constrains and can be obtained through the published article

PAPER III: COMPREHENSIVE ANALYSIS OF TRANSCRIPTOME VARIATION UNCOVERS KNOWN AND NOVEL DRIVER EVENTS IN T-CELL ACUTE LYMPHOBLASTIC LEUKEMIA

Zeynep Kalender Atak[*1], Valentina Gianfelici[*2,3], Gert Hulselmans[*1], Kim De Keersmaecker[*2], Arun George Devasia[1,2], Ellen Geerdens[2], Nicole Mentens[2], Sabina Chiaretti[3], Kaat Durinck[4], Anne Uyttebroeck[5], Peter Vandenberghe[2], Iwona Wlodarska[2], Jacqueline Cloos[6], Robin Foà[3], Frank Speleman[4], Jan Cools[2,#], and Stein Aerts[1,#]

[1] Laboratory of Computational Biology, Center for Human Genetics, KU Leuven, Leuven, Belgium

[2] Laboratory for the Molecular Biology of Leukemia, Center for Human Genetics, KU Leuven and Center for the Biology of Disease, VIB, Leuven, Belgium

[3] Division of Hematology, Department of Cellular Biotechnologies and Hematology, 'Sapienza' University of Rome, Italy

[4] Center for Medical Genetics, Ghent University, Ghent, Belgium

[5] Pediatric Hemato-Oncology, University Hospitals Leuven, Leuven, Belgium

[6] Pediatric Oncology/Hematology and Hematology, VU Medical Center, Amsterdam, The Netherlands

* equal contribution

Manuscript is accepted for publication in *PLoS* Genetics.

ABSTRACT

RNA-seq is a promising technology to re-sequence protein coding genes for the identification of single nucleotide variants (SNV), while simultaneously obtaining information on structural variations and gene expression perturbations. We asked whether RNA-seq is suitable for the detection of driver mutations in T-cell acute lymphoblastic leukemia (T-ALL). These leukemias are caused by a combination of gene fusions, over-expression of transcription factors and cooperative point mutations in oncogenes and tumor suppressor genes. We analyzed 31 T-ALL patient samples and 18 T-ALL cell lines by high-coverage paired-end RNA-seq. First, we optimized the detection of SNVs in RNA-seq data by comparing the

results with exome re-sequencing data. We identified known driver genes with recurrent protein altering variations, as well as several new candidates including *H3F3A, PTK2B*, and *STAT5B*. Next, we determined accurate gene expression levels from the RNA-seq data through normalizations and batch effect removal, and used these to classify patients into T-ALL subtypes. Finally, we detected gene fusions, of which several can explain the over-expression of key driver genes such as *TLX1, PLAG1, LMO1,* or *NKX2-1*; and others result in novel fusion transcripts encoding activated kinases (*SSBP2-FER* and *TPM3-JAK2)* or involving *MLLT10*. In conclusion, we present novel analysis pipelines for variant calling, variant filtering, and expression normalization on RNA-seq data, and successfully applied these for the detection of translocations, point mutations, INDELs, exon-skipping events, and expression perturbations in T-ALL.

AUTHOR SUMMARY

The quest for somatic mutations underlying oncogenic processes is a central theme in today's cancer research. High-throughput genomics approaches including amplicon re-sequencing, exome re-sequencing, full genome re-sequencing, and SNP arrays have contributed to cataloguing driver genes across cancer types. Thus far transcriptome sequencing by RNA-seq has been mainly used for the detection of fusion genes, while few studies have assessed its value for the combined detection of SNPs, INDELs, fusions, gene expression changes, and alternative transcript events. Here we apply RNA-seq to 49 T-ALL samples and perform a critical assessment of the bioinformatics pipelines and filters to identify each type of aberration. By comparing to exome re-sequencing, and by exploiting the catalogues of known cancer drivers, we identified many known and several novel driver genes in T-ALL. We also determined an optimal normalization strategy to obtain accurate gene expression levels and used these to identify over-expressed transcription factors that characterize different T-ALL subtypes. Finally, by PCR, cloning, and *in vitro* cellular assays we uncover new fusion genes that have consequences at the level of gene expression, oncogenic chimaeras, and tumor suppressor inactivation. In conclusion, we present the first RNA-seq data set across T-ALL patients and identify new driver events.

INTRODUCTION

T-cell acute lymphoblastic leukemia (T-ALL) is an aggressive malignancy that accounts for approximately 15% of pediatric and 25% of adult ALL cases. Despite improved outcome over the years, about 25% of children and 50% of adults still fail to respond to intensive chemotherapy protocols or relapse [1]. Improved understanding of T-ALL biology through the identification and characterization of oncogenic lesions is expected to lead to a better prognostic classification and the development of new targeted therapeutic strategies.

T-ALL is caused by the accumulation of multiple oncogenic mutations that have been identified through characterization of chromosomal aberrations and candidate

gene sequencing [2]. Chromosomal translocations in T-ALL frequently involve the T-cell receptor (*TCR*) loci, whereby *TCR* regulatory elements become juxtaposed to genes that are normally not expressed in T-cells [3,4]. In this way, a specific set of recurrently over-expressed TFs have been documented, including *TLX1, TLX3, TAL1, LMO1, HOXA,* and *NKX* family members [5]. T-ALL samples expressing each of these transcription factors show a distinctive gene expression signature and as such these transcription factors define distinct molecular subtypes in T-ALL [6]. Chromosomal rearrangements can also lead to large chromosomal deletions and amplifications; to focal gene deletions or amplifications, such as *CDKN2A* deletion and *MYB* duplication [7,8]; and to in-frame fusion genes encoding chimeric proteins with oncogenic properties such as the constitutively active *NUP214-ABL1* fusion kinase [9]. In addition, point mutations and small INDELs have also been described leading to oncogenic events, such as mutations activating *NOTCH1* that occur in more than 60% of T-ALL cases [10], or mutations in cytokine receptors and tyrosine kinases such as *IL7R* and *JAK3* [11-17]. The latter may lead to new opportunities for molecularly tailored therapies with kinase inhibitors [12,16,18,19].

With the advent of next generation sequencing (NGS) technologies, our sequencing capacity has significantly improved in the past five years. It is now possible to apply targeted re-sequencing, exome sequencing (Exome-seq), whole genome sequencing (WGS), whole transcriptome sequencing (RNA-seq) or a combination of these, to investigate individual genomes, especially those related to disease [20]. Also for T-ALL, these NGS approaches have recently proven their value in the discovery of novel driver genes [13,14,17,21]. We previously identified a spectrum of new oncogenic driver genes using Exome-seq on 67 T-ALLs, and described clear differences between pediatric and adult patients [17]. In particular, we identified *CNOT3* as a tumor suppressor mutated in 8% of adult T-ALL cases and mutations affecting the ribosomal proteins *RPL5* and *RPL10* in 10% of pediatric T-ALLs [17]. Similarly, whole genome sequencing of early T-cell precursor ALL cases led to the identification of mutations in several new oncogenes and tumor suppressor genes affecting cytokine signaling, T-cell development and histone-modifying genes [2,13]. However, the potential of RNA-seq for the discovery of driver genes in T-ALL remains unexplored.

In the present study, we applied paired-end RNA-seq on 49 T-ALL samples (31 patients, 18 cell lines) to gain insights in the transcriptome landscape of T-ALL. First, we show that identification of somatic single nucleotide variants (SNV) and recurrently mutated driver genes is feasible on RNA-seq data, even without matched normal samples (e.g., germline or remission DNA). We identify *STAT5B, H3F3A*, and *PTK2B* as candidate cancer genes in T-ALL. This becomes possible when (1) optimal read mapping and SNV calling procedures are applied; and (2) functional annotation, gene expression, or additional sequencing data from other cohorts is used to prioritize the true driver genes. Next, we optimized gene expression measurements using multiple normalization strategies, and showed that classical gene expression studies (e.g., clustering) are feasible on normalized RNA-seq data.

We also detected new fusion genes (*SSBP2-FER* and *TPM3-JAK2*) and used gene expression data to determine the consequence of observed chromosomal rearrangements on the over-expression of key driver genes. Finally, we searched for significant alternative transcript events (ATE) but besides one coherent exon-skipping event in *SUZ12*, we found relatively few candidate ATEs in T-ALL. In conclusion, through a combination of the analysis of gene expression levels, fusion transcripts, SNVs, and INDELs, we could identify known and new driver alterations in T-ALLs and novel potential targets for therapy.

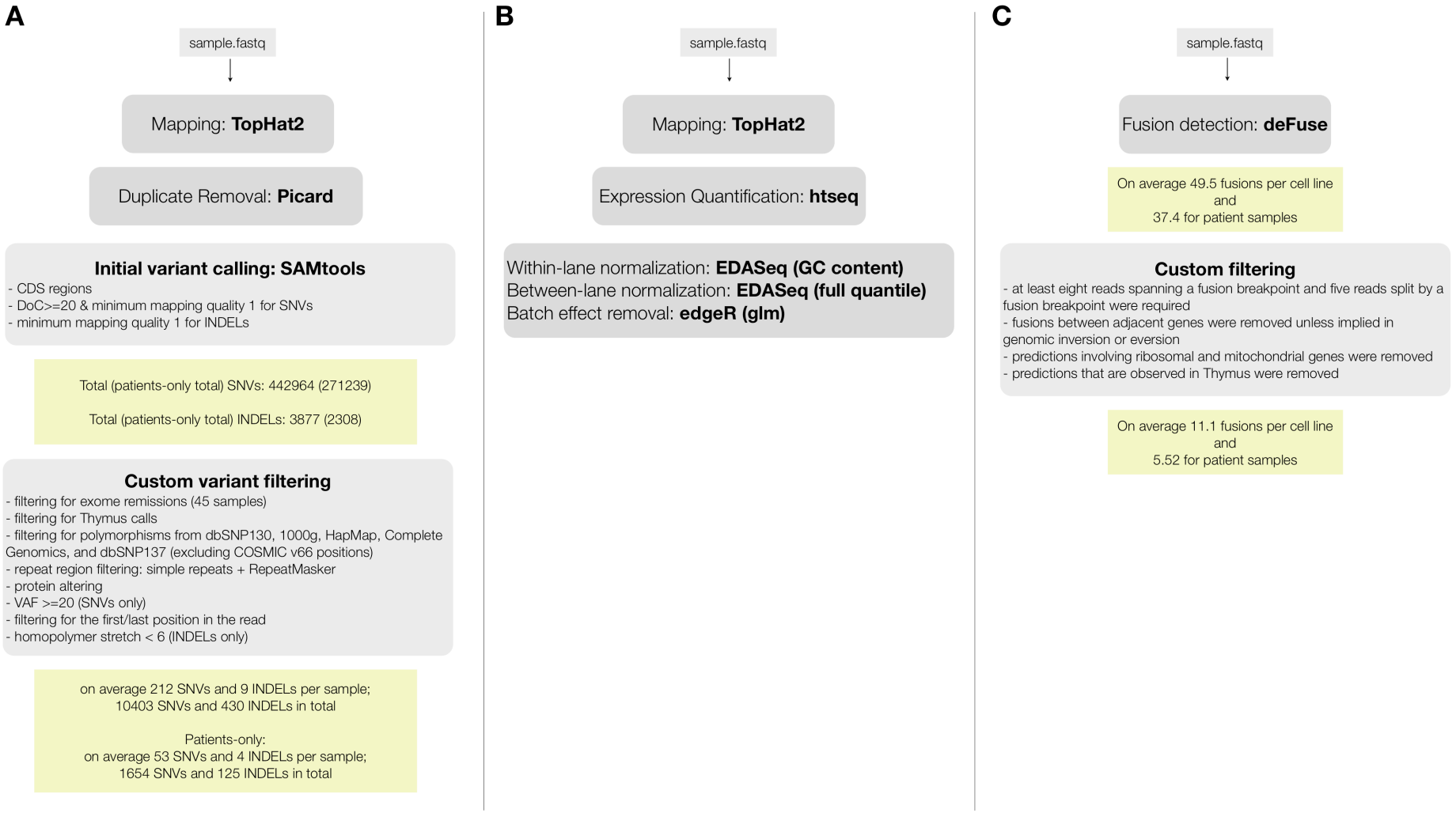
RESULTS

Correct SNV and INDEL calling on RNA-seq data depends on accurate read mapping

We performed paired-end RNA-seq on 31 T-ALL patients, 18 T-ALL cell lines, and 1 normal thymus sample. We obtained on average ~110 million reads per sample, leading to an average coverage of ~88X (**Table S1.A**). To assess the quality of detecting SNVs from the RNA-seq data, we compared the RNA-seq to Exome-seq data. For 16/18 of the cell lines and for 20/31 patient samples we had exome data available (previously generated [17] or obtained for this study, **Table S2**). For the exome data analysis, we followed the pipeline of mapping, SNV and somatic mutation detection that we validated previously [17] (using BWA, GATK, SomaticSniper, and Variant Effect Predictor (VEP)) [22-25]. For the RNA-seq data we used TopHat2 [1,26] for mapping, SAMTools [2,27] for SNV detection, and VEP [3,4,25] for variant annotation (**Figure 1.A**).

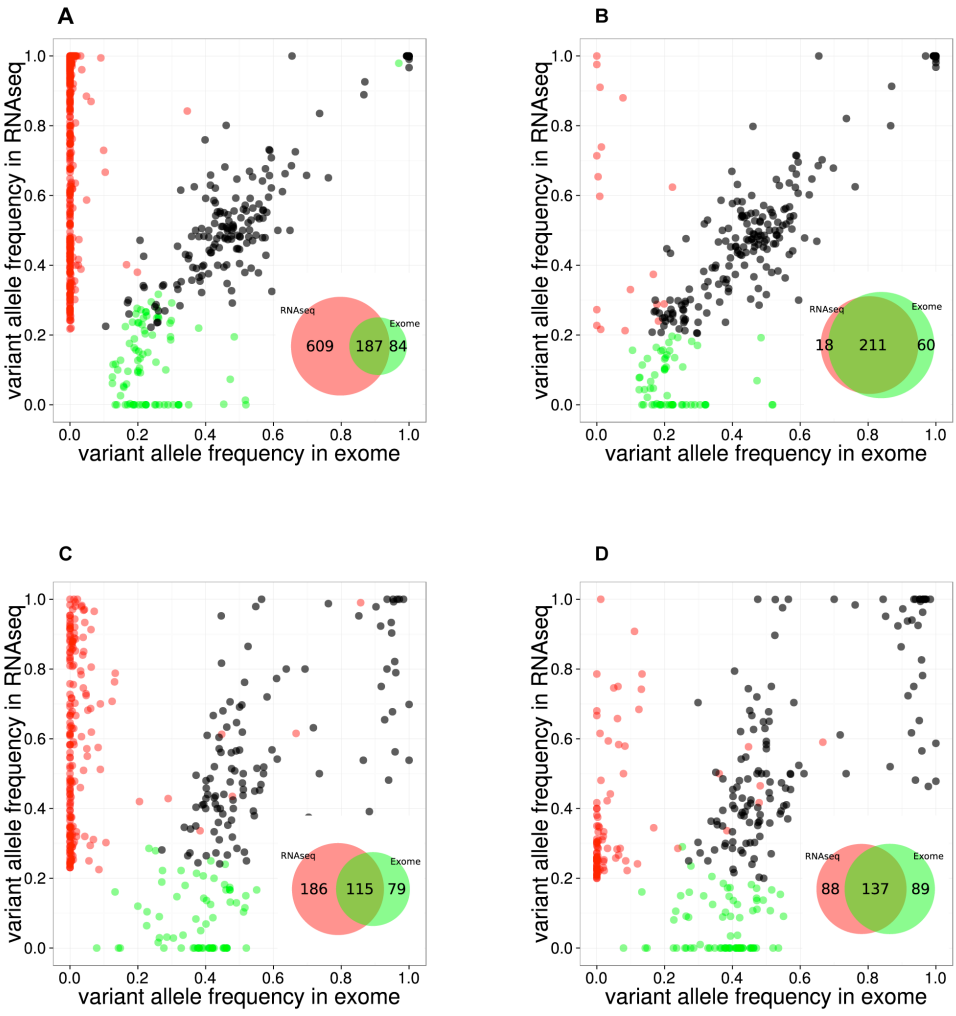
By comparing positions that had a coverage of at least 20x in both RNA-seq and Exome-seq, combined with Sanger re-sequencing of a subset of positions, we found that the accuracy of SNV calling in RNA-seq strongly depends on the read mapping, corroborating earlier observations [5,28,29] (**Figure S1**). We found that mapping RNA-seq reads to the genome (as used by TopHat version 1.3.3) is prone to errors when dealing with paralogous genes, as observed by the prediction of false positive SNVs in *KIF4A* and *GLUD1* due to erroneous mapping to *KIF4B* and *GLUD2* (both pseudogenes with no introns) (**Figure S1**). However, these errors were resolved by mapping to the transcriptome. In the case of the RPMI8402 cell line, 877 SNVs were found by mapping to the genome, while this number was reduced to 283 SNVs when mapping to the transcriptome. Mapping to the transcriptome did not only reduce the number of RNA-seq exclusive calls but also increased the overlap with the Exome-seq calls (**Figure 2, Figure S2**).

Figure 1. RNA-seq data analysis pipelines for (A) variant calling and filtering to detect point mutations, (B) fusion detection and annotation, (C) gene expression analysis.



However, transcriptome mapping also has limitations as it relies on current gene and isoform annotation. We observed that a combination of transcriptome and genome mapping provides the best solution. It is important that all reads are mapped twice to the genome, independently of each other; once as entire read and once as split read. This has become possible in TopHat2 by setting the option "*read-realign-edit-dist*" to zero. Our analysis reveals that this mapping approach results in the best overlap of SNVs compared to exomes (**Figure 2, Figure S3**). This mapping strategy not only improves the alignment accuracy by preventing misalignment to pseudogenes, but also leads to identification of the most likely isoform structure of a gene by mapping the reads independently both to the transcriptome and to the genome and then selecting the best possible alignment.

Figure 2. Comparison between RNA-seq and exome-seq.
Variant Allele Frequency plots for evaluating two RNA-seq mapping strategies for two example samples, namely the RPMI8402 cell line (**A**, **B**) and the TLE79 patient sample (**C**, **D**). On the left are the results of mapping with TopHat 1.3.3. (**A**,**C**), while on the right are the results of mapping with TopHat 2.0.5 with forced re-mapping of all reads to the genome. The SNVs that have at least 20 reads in exome-seq and RNA-seq are plotted. Red and green dots represent the SNVs that are detected only in RNA-seq and only in exome-seq, respectively, while black dots represent the SNVs that are called in both. Venn diagrams are produced from the points represented in the graphs.



Using the optimized mapping and filtering strategy we identified 436974 SNVs across 49 samples. By using samples for which both the exome and the transcriptome were sequenced several aspects of SNV detection in RNA-seq data can be evaluated, such as sensitivity, specificity, and allelic imbalance. Regarding sensitivity, we found that on average, 32% of the SNVs that are called in Exome-seq were also called by the RNA-seq (**Table S3**). Similar ratios were observed when comparing validated somatic SNVs from Exome-seq/WGS to RNA-seq SNVs: 36% in a triple negative breast cancer study [6,30], and 41% in a lymphoma study [7,8,31]. We observed that the sensitivity varies considerably between samples, and strongly correlates with the average depth of coverage of the sample (**Figure S4**).

Regarding specificity, we found that the remaining RNA-seq-only and exome-only SNVs (for positions where both have at least 20x coverage) are found mainly with a low variant allele frequency (VAF) and are therefore likely due to arbitrary VAF and coverage thresholds. For example, on the RPMI8402 and TLE79 samples, many RNA-seq-only SNVs (9/18 and 61/88 respectively) have a VAF below 40%. Regarding allelic imbalance, we found that of all heterozygous Exome SNVs with more than 20X coverage, the majority (2914/4043 or 72%) were also heterozygous SNVs in RNA-seq. Of the remaining SNVs, many (988/4043) are homozygous reference in the RNA-seq (i.e., not detected). A small fraction we can almost certainly attribute to allelic imbalance, namely the 141/4043 SNVs (3.5%) that are homozygous variant in the RNA-seq, indicating that for those only the variant allele is expressed (or the gene is only expressed in cancer cells that harbor the variant).
Next we asked whether small insertions and deletions (INDELs) can be detected from RNA-seq data. As with the SNVs, we used the Exome-seq data for assessing the quality of our INDEL detection strategy. On average, 47.5% of the INDELs that were detected by RNA-seq were also found in the Exome-seq (unfiltered) INDEL calls. However, only 4% of the Exome-seq INDELs (for which the region containing INDEL is covered by at least 3 reads in RNAseq data) were found back in the RNA-seq calls (**Table S3**). To investigate this sensitivity issue, we evaluated ten validated INDELs that we previously identified with Exome-Seq [9,17](**Table S4**). Three of the ten INDELs were also identified in the RNA-seq data using the default SAMTools parameters (see Materials and Methods). Of the seven missed INDELS, two are found in a gene that is not expressed; another two are clearly present in the RNA-seq data when inspected manually with IGV, but did not reach the default threshold (see Materials and Methods); and the last three are effectively discordant between RNA-seq and Exome-seq, as they show only reads with reference sequence (**Figure S5**). Re-mapping of the reads with BWA [10,22] on the transcriptome followed by BLAT [11-17,32] on the genome improved the INDEL identification, now revealing the *KDM6A* INDEL in TLE87 and *PTEN* INDEL in TLE92, which were previously missed (**Figure S6.A-B**). It is notable that the combination of TopHat2 (to transcriptome only) and BLAT does not correctly detect these two INDELS (**Figure S6.C-D**). We conclude that INDEL detection on RNA-seq data is feasible,

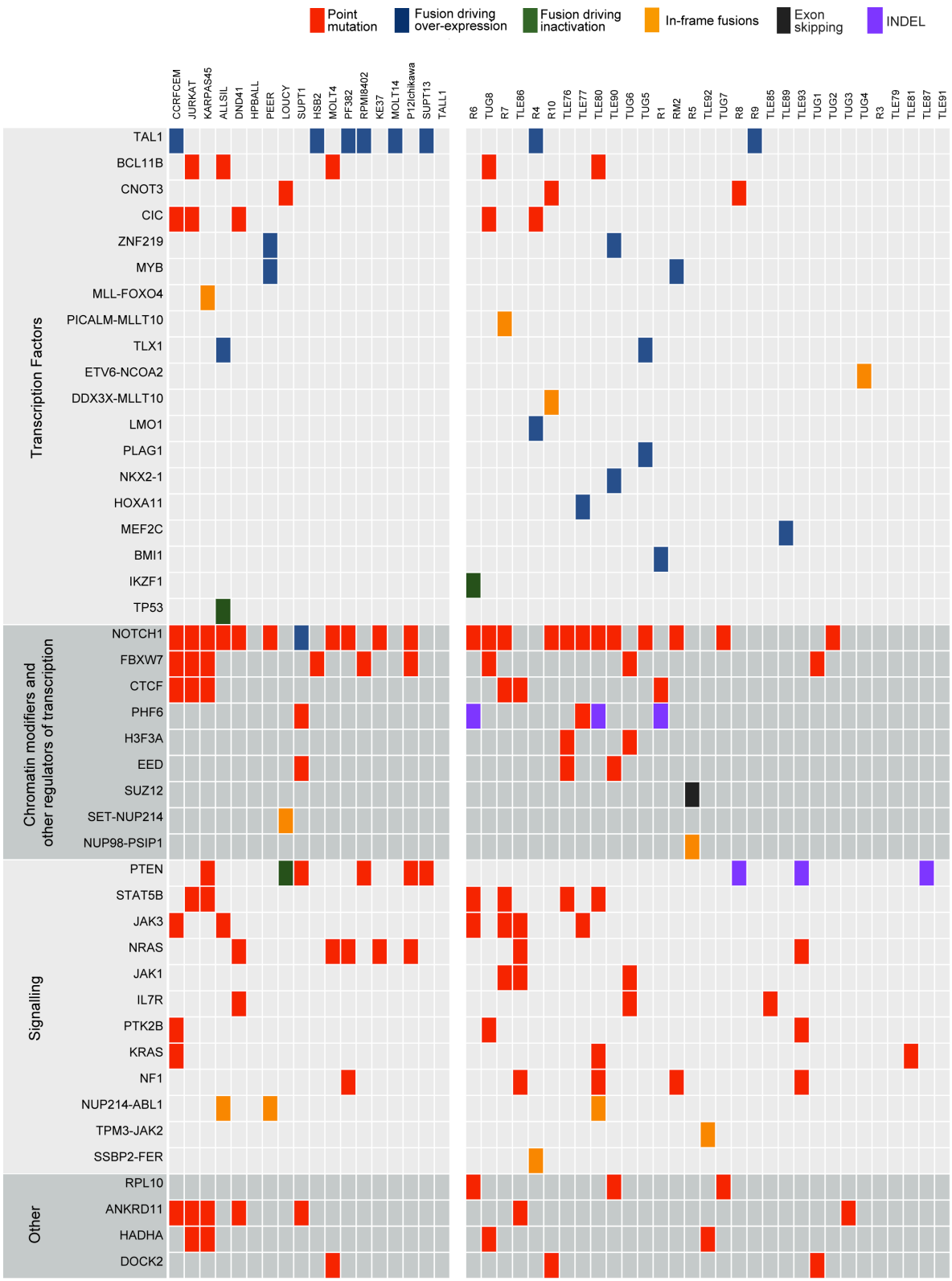
yet technically challenging, and that the fraction of INDELs compared to SNVs is moderate (see also the next Section and Figure 3 below).

Leveraging diagnosis-only RNA-seq data with the T-ALL body of knowledge to identify mutated cancer genes

Our next aim was to select candidate driver genes using the collected SNVs and INDELS. To remove germline variants we initially removed all SNPs present in dbSNP [12,16,18,19,33], 1000genomes [20,34], the Complete Genomics genomes [13,14,17,21,35], and those detected in our own exome data from normal samples (39 from our earlier work [17] and 6 from this study). We, however, retained those variants also present in the COSMIC [17,36] database, since SNP databases are know to contain also some disease-specific SNVs. Some examples of SNVs that are likely driver mutations, but that are also present in polymorphism databases are: *JAK3* A572V in R7, and *FBXW7* R425C in TUG1. With this filtering, we obtained a final list of 10403 protein-altering SNVs and 430 protein-altering INDELs, with a median of 63 SNVs and 4 INDELS per sample (**Table S1.B**). Cell lines harbored significantly more mutations than patient samples (Mann-Whitney test p-value=1.095e-05), as previously also observed by exome-seq [2,13,17].

As a first approach to identify candidate T-ALL driver genes, we selected all genes that contained a protein-altering mutation in at least two of the 31 patient samples (for recurrence we did not take cell lines into account). This process resulted in the selection of 213 genes (**Table S5**). We found that this list is strongly enriched for genes related to T-ALL and to cancer in general, with "precursor T-cell lymphoblastic leukemia-lymphoma" as the most highly enriched function (p-value 1.35E-11 by Ingenuity Pathway Analysis) (**Table S6**). The list of 213 candidates contained many known T-ALL driver genes (**Figure 3**), such as *NOTCH1, BCL11B, FBXW7, IL7R, JAK1* and *JAK3*; and it also contained the drivers *CNOT3* and *RPL10*, recently identified in our exome re-sequencing study [17]; and *CTCF*, which was recently reported to be recurrently mutated in ETP-ALL [13,17]. In addition, the candidate list contained two established cancer driver genes involved in other cancer types, but not yet reported to be mutated in T-ALL, namely *H3F3A* and *CIC*. These genes were reported recently by Vogelstein [22-25,37] to be true cancer drivers.

Figure 3. Point mutations and gene fusions organized into functional categories (next-page).
Protein altering mutations and INDELs, alternative splicing events and validated fusions are shown. Red boxes indicate protein-altering mutations (i.e. nonsense, missense and splice site mutations), purple boxes indicate frameshift INDELs whereas blue, green and orange boxes represent fusion events resulting in over-expression of the partner gene, inactivation of the partner gene or generation of a chimeric protein, respectively, and finally black boxes indicating alternative splicing events.


Point mutation
Fusion driving over-expression
Fusion driving inactivation
In-frame fusions
Exon skipping
INDEL
CCRFCEM
JURKAT
KARPAS45
ALLSIL
DND41
HPBALL
PEER
LOUCY
SUPT1
HSB2
MOLT4
PF382
RPMI8402
KE37
MOLT14
P12Ichikawa
SUPT13
TALL1
R6
TUG8
R7
TLE86
R4
R10
TLE76
TLE77
TLE80
TLE90
TUG6
TUG5
R1
RM2
R5
TLE92
TUG7
R8
R9
TLE85
TLE89
TLE93
TUG1
TUG2
TUG3
TUG4
R3
TLE79
TLE81
TLE87
TLE91
Transcription Factors
TAL1
BCL11B
CNOT3
CIC
ZNF219
MYB
MLL-FOXO4
PICALM-MLLT10
TLX1
ETV6-NCOA2
DDX3X-MLLT10
LMO1
PLAG1
NKX2-1
HOXA11
MEF2C
BMI1
IKZF1
TP53
Chromatin modifiers and other regulators of transcription
NOTCH1
FBXW7
CTCF
PHF6
H3F3A
EED
SUZ12
SET-NUP214
NUP98-PSIP1
Signalling
PTEN
STAT5B
JAK3
NRAS
JAK1
IL7R
PTK2B
KRAS
NF1
NUP214-ABL1
TPM3-JAK2
SSBP2-FER
Other
RPL10
ANKRD11
HADHA
DOCK2

We identified two patient samples (TLE76 and TUG6) with *H3F3A* mutations both on the K28 residue that is a mutational hotspot in glioblastoma [38]. This mutation was confirmed somatic in the TUG6 sample. Sequencing of this hotspot in additional T-ALL samples indicated a low frequency of *H3F3A* K28 mutation in T-ALL (detected in 3 of 102 cases).

Next we asked if we could identify additional genes in the candidate list that could be linked to T-ALL. We wanted to utilize the genes that are known to be involved in T-ALL as a guide for identifying additional candidates. To this end we used our gene prioritization approach ENDEAVOUR [39], which scores candidate genes based on a set of training genes. It builds a profile based on the training genes (integrating information on protein-protein interactions, genetic interactions, gene expression, text-mining, sequence homology, Gene Ontology, and protein domains) and then prioritizes the candidate genes for their similarity to the derived profile. As training set we used all known drivers, and as test set we used all the 213 candidates with at least two patient mutations (excluding the genes that are in the training set). We reasoned that this would reveal the genes with strong similarity to the known drivers and such genes would be good candidate drivers. We found 45 significantly ranked genes with two interesting genes at the top of the ranking, namely *PTK2B* and *STAT5B* that are involved in JAK/STAT signaling (**Table S7**). Furthermore, the list contained genes for which we had identified single T-ALL cases with a somatic mutation in our previous exome study: *ANKRD11*, *CTCF*, *DOCK2*, *H3F3A,* and *HADHA*. We did not select these genes before in our Exome-seq cohort [17] because they were only mutated in one of the 39 samples we analyzed. Now, with the RNA-seq cohort, we thus found additional samples with mutations in these genes.

Optimized gene expression measurements and batch effect removal from RNA-seq data identify co-expression modules and T-ALL subtypes

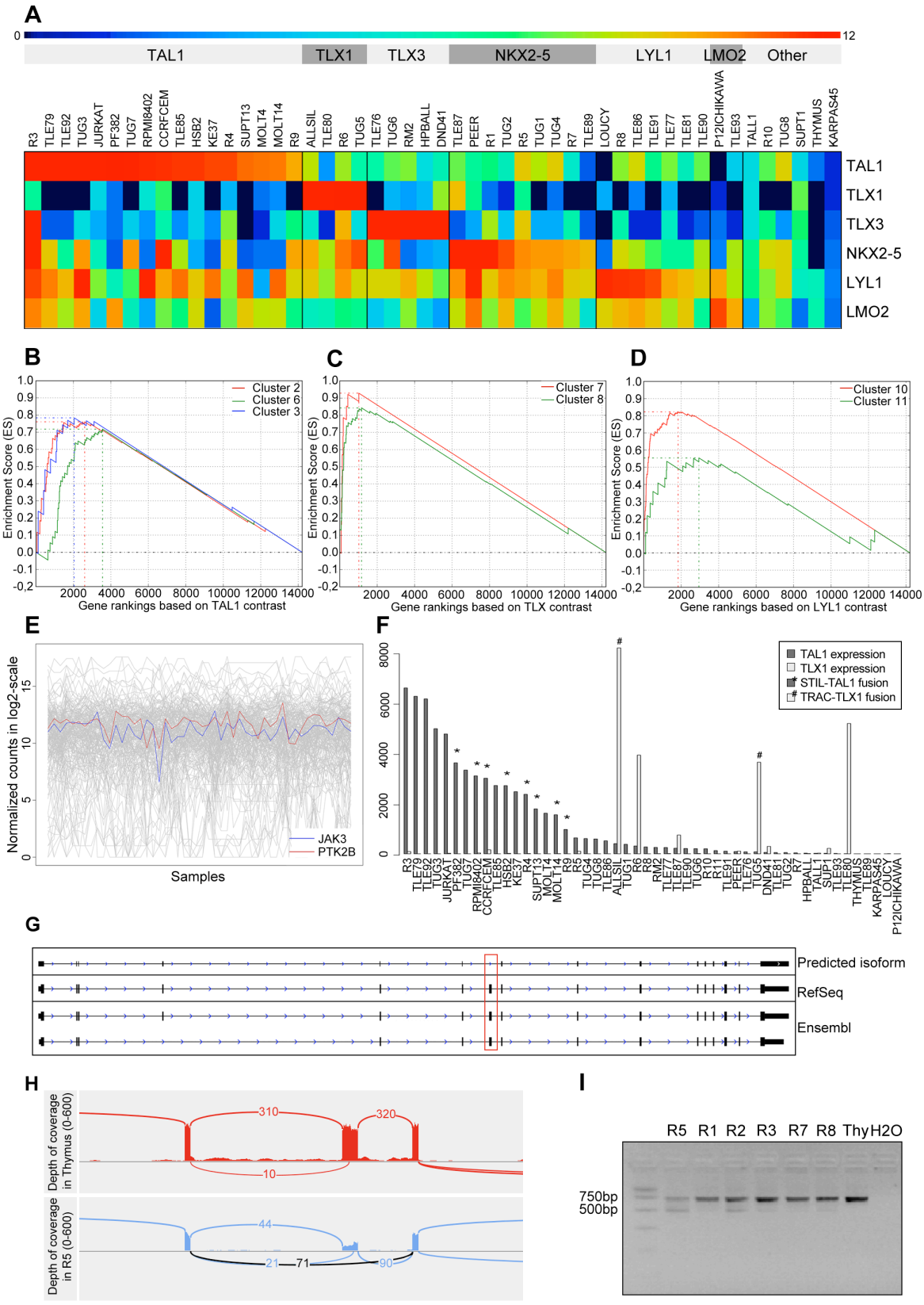
T-ALL is characterized by the overexpression of transcription factors (TFs), such as *TLX1, TLX3, TAL1*, and the *HOXA* family members [6]. Therefore, identifying and analyzing expression perturbations in a T-ALL cohort is highly relevant. To obtain accurate gene expression levels from the mapped RNA-seq reads, we followed the procedure outlined in **Figure 1.B**, including read aggregation, GC-normalization, length normalization, and between-sample normalization (see Materials and Methods). In addition, we removed a batch effect that was clearly present in the data set using a Generalized Linear Model (GLM, see Materials and Methods) (**Figure S7**). It is notable that transcript-based expression analysis conducted with *cufflinks* revealed the same batch effect linked to the origin of the sample, thereby confirming a technical bias in the data set (**Figure S7.B**, see Materials and Methods).

We next looked at the expression values of *TLX1*, *TLX3*, *TAL1*, and other important TFs in T-ALL. Clustering of *TLX1, TLX3*, and *TAL1* expressing samples confirmed that the correct samples (based on karyotyping and molecular

analysis) showed over-expression of the respective TF (**Figure 4.A**). Indeed, 8 samples that harbored a *STIL-TAL1* rearrangement showed high *TAL1* expression (**Figure 4.D**). Note that also other samples with high *TAL1* expression were detected. This fits with a previously reported observation of *TAL1* over-expression in the absence of a translocation in T-ALL [6,40].

To assess the accuracy of our expression values obtained after normalization, batch effect removal and clustering, we tested whether previously published gene signatures associated with *TAL1*, *TLX* (*TLX1* and *TLX3*) and *LYL1* can be detected also in our data set [41]. We used 13 gene signatures obtained by Soulier *et al* using a microarray study on 92 primary T-ALL samples [41]. Gene set enrichment analysis shows that our *TAL1* expressing cases are significantly associated with *TAL1* signatures, whereas our *TLX* over-expressing cases are associated with the *TLX* signature (7, 8) and the *LYL1* cases with the *LYL1* signature (10, 11). This analysis confirms that the obtained expression data represent meaningful values and sample clustering produces gene lists that are biologically meaningful (**Figure 4.B**).We next used the gene expression information as a guide to assist in the detection of relevant mutations. We found that the expression profile of *PTK2B*, a candidate driver identified above by ENDEAVOUR, significantly correlated with the *JAK3* expression profile (PTM, with p-value threshold at 1E-5, see Materials and Methods) (**Figure 4.C**). Indeed, *PTK2B* was previously implicated in *IL-2* mediated signaling and JAK/STAT signaling, and was shown to physically interact with *JAK3* [42]. These data warrant further investigation of *PTK2B* as an important tyrosine kinase in T-ALL case with activated JAK/STAT signaling.

Figure 4. Validation and discovery using gene expression data, and SUZ12 ATE (next page).
(**A**) Classification of the samples using the TFs that are known to be overexpressed in T-ALL. Using the expression patterns of *TAL1*, *TLX1*, *TLX3*, *NKX2-5*, *LYL1* and *LMO2* we could discriminate the samples in to six distinct clusters. The heatmap is plotted with the normalized log2(count) values. Gene set enrichment analysis curves are displayed for (**B**) enrichment of *TAL1* associated clusters 2, 6 and 3 in *TAL1* based ranking, (**C**) enrichment of *TLX* associated clusters 7 and 8 in *TLX* based ranking, and (**D**) enrichment of *LYL1* associated clusters 10 and 11 in *LYL1* based ranking of the genes. (**E**) Expression of *JAK3* and *PTK2B* across samples is significantly correlated (with PTM p-value 1E-5). (**F**) Normalized expression values of *TAL1* and *TLX1* with translocations affecting these genes indicated. The samples with a translocation have elevated expression of the affected gene, showing the driver potential of the fusion event. There are additional samples with high expression of *TLX1* and *TAL1* without the indicated fusions, pointing to other mechanisms of activating these genes. (G) Predicted SUZ12 transcript aligned with the known *SUZ12* isoforms. Dotted red box indicates the location of the exon-skipping event. (H) The sashimi plot below shows the junction (in black) supporting the exon-skipping event in patient sample R5 with respect to Thymus.(I) Agarose gel electrophoresis of the RT-PCR products for validation of *SUZ12* exon skipping event. The two isoforms are clearly detected in R5 and to a minor extent in the other T-ALL samples while Thymus shows only the canonical transcript.


A
0
12
TAL1
TLX1
TLX3
NKX2-5
LYL1
LMO2
Other
R3 TLE79 TLE92 TUG3 JURKAT PF382 TUG7 RPMI8402 CCRFCEM TLE85 HSB2 KE37 R4 SUPT13 MOLT4 MOLT14 R9 ALLSIL TLE80 R6 TUG5 TLE76 TUG6 RM2 HPBALL DND41 TLE87 PEER R1 TUG2 R5 TUG1 TUG4 R7 TLE89 LOUCY R8 TLE86 TLE91 TLE77 TLE81 TLE90 P12ICHIKAWA TLE93 TALL1 R10 TUG8 SUPT1 THYMUS KARPAS45
TAL1
TLX1
TLX3
NKX2-5
LYL1
LMO2
B
Cluster 2
Cluster 6
Cluster 3
Enrichment Score (ES)
Gene rankings based on TAL1 contrast
C
Cluster 7
Cluster 8
Enrichment Score (ES)
Gene rankings based on TLX contrast
D
Cluster 10
Cluster 11
Enrichment Score (ES)
Gene rankings based on LYL1 contrast
E
Normalized counts in log2-scale
JAK3
PTK2B
Samples
F
TAL1 expression
TLX1 expression
STIL-TAL1 fusion
TRAC-TLX1 fusion
G
Predicted isoform
RefSeq
Ensembl
H
Depth of coverage in Thymus (0-600)
310
320
10
Depth of coverage in R5 (0-600)
44
21
71
90
I
R5 R1 R2 R3 R7 R8 ThyH2O
750bp
500bp

T-ALL presents robust transcript isoform usage

To our knowledge, only very few cancer specific alternative transcript events (ATE) have been described for any cancer type [43-45], and no ATE is reported for T-ALL. In contrast to SNVs, INDELS, CNAs, and fusions, which are all curated and present in large numbers in public cancer mutation databases (e.g., COSMIC [36] , CENSUS [46]), we could not find driver ATEs in those databases (although splice sites represent an important class of cancer mutations). If ATEs represent an important, yet underestimated, type of somatic variation in cancer, we would expect at least some of the known cancer driver genes to present a significant ATE. We thus asked whether novel variations could be found in these genes in the form of ATEs.

To this end, we applied *cufflinks* and *cuffdiff* (see Materials and Methods) and found significant ATEs in 12 of the 47 known driver genes (*BCL11B, FLT3, IL7R, LCK, MYB, NKX2-1, SFTA3, RPL10, RUNX1, SETD2, SUZ12, and TAL1) (**Table S8**). However, when we manually inspected* these events in IGV, we found only two interesting cases. One case represents an unambiguous skipping of exon 7 in *SUZ12*, occurring in several patient samples, but most significant (Cuffdiff p-value 5.10^{-5}) in the R5 patient sample, and absent in thymus (**Figure 4.E**). and a potential, but less clear, skipping of exon 8 in *LCK* in three samples (**Figure S8**). Exon 7 of *SUZ12* is a canonical exon (present in all known isoforms) according to RefSeq, Ensembl, and UCSC annotation. The ATE we observe is a heterozygous event with the wild-type junction supported by 90 reads and the novel junction supported by 71 reads. RT-PCR clearly confirmed the exon-skipping event in R5 and to a minor extent in other samples, while being absent in the thymus (Fig 4.F). The functional consequences of these splice variants remain to be determined, but the fact that these variants are both in-frame suggests that these proteins could be functional protein isoforms (**Figure S8 and S9**). Overall, relatively few significant ATEs are detected, and no obvious ATEs are found with consequences on the protein structure, therefore T-ALL presents robust isoform usage at the current resolution of sequencing and analysis.

Detection and validation of known and novel fusion transcripts

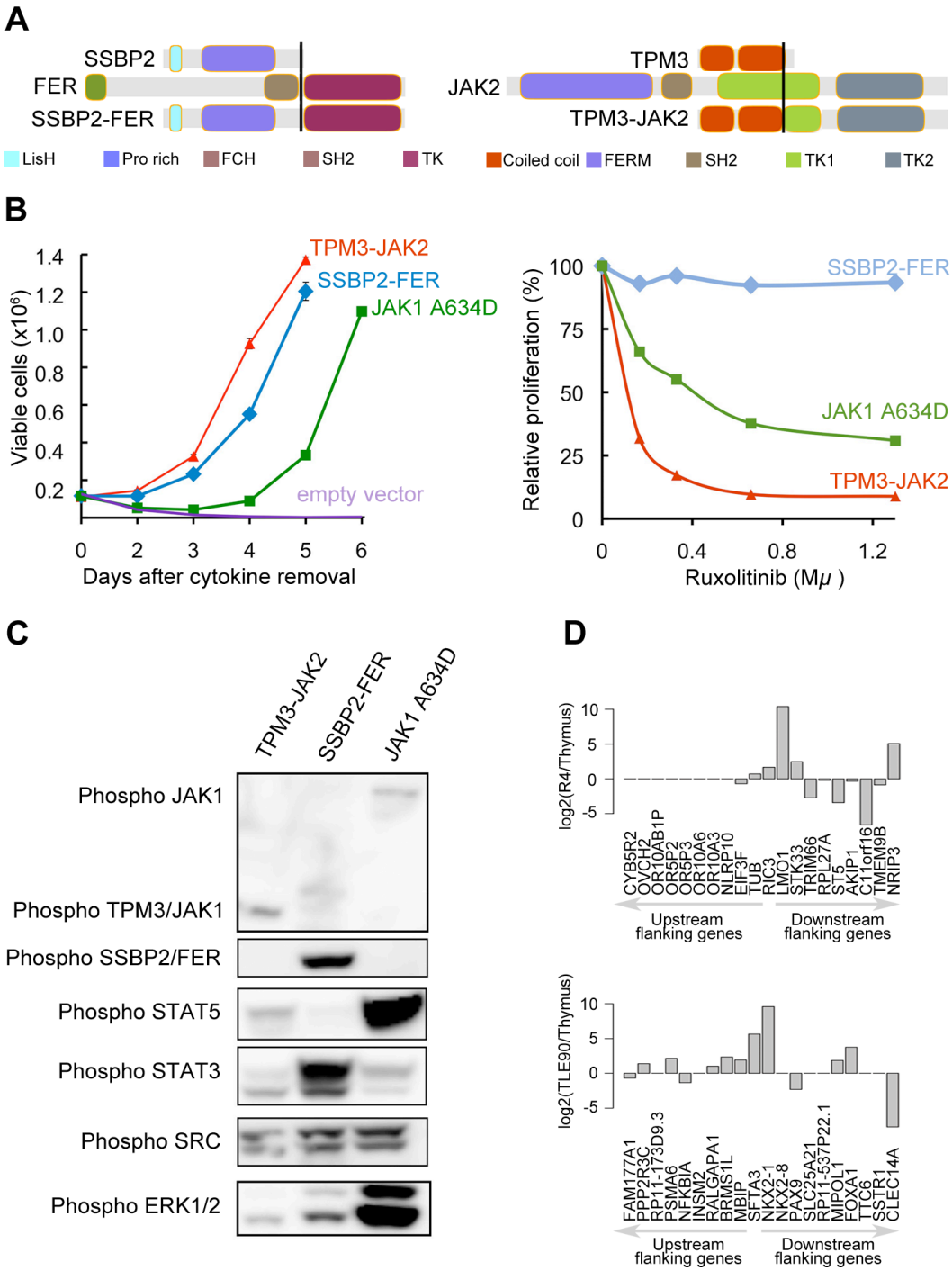
Most of the T-ALL cases harbor chromosomal rearrangements that lead to the generation of fusion genes or ectopic expression of genes due to juxtaposition to strong promoters or regulatory sequences. Chromosomal translocations involving the TCR genes are largely underestimated by karyotyping and the TCR partner genes remained unidentified in several cases [4,47]. On the other hand, a multitude of mechanisms other than translocations could cause ectopic expression of oncogenes [48]. To detect fusion transcripts, we used the defuse algorithm on our entire dataset [49]. Briefly, this method identifies candidate gene fusions by discordant alignments produced by spanning reads (each read in the read pair aligns to a different gene) and by split reads (reads that harbor a fusion boundary). The total number of predicted fusions initially was 1,160 and 1,265 in patient and cell line samples,

respectively. Also in normal thymus RNA, 60 fusion transcripts were detected. Next, we implemented additional filters, considering only predictions supported by 8 or more spanning reads and 5 or more split reads. Furthermore, we removed fusions involving ribosomal genes, mitochondrial genes and fusions between adjacent genes, as these could be caused by read-through or trans-splicing [50,51] (**Figure 1.C**).
After applying these filters, we obtained an average of 5.5 fusion events per patient sample and 11.1 per cell line (**Table S1.C**). In total, 397 candidate genes are involved as potential partner in a gene fusion (**Table S9**). Details on the fusion breakpoints and validation of the novel candidate fusion transcripts are reported in Tables S9 and S12 (see also Materials and Methods: RT-PCR and Sanger Sequencing).

First, to determine the relevance of these predicted fusion transcripts we looked at functional enrichment of these genes. 278 of 397 genes correspond to functionally annotated protein-coding genes according to DAVID functional enrichment [52,53]. Furthermore, this set is strongly enriched for cancer-related genes, and more specifically for genes involved in AML (p value 4.48E-10) and T-ALL (p value 4.47E-5), including *TP53, STAT5B, NOTCH1, IL7R, IKZF1, CDKN2A, MLLT10, ETV6,* and *ABL1*.

Second, we specifically analyzed the 27 in-frame fusions, predicted to encode chimeric proteins (**Table S10**). This list contained known oncogenic fusion genes, including *NUP214-ABL1* (n=2), *MLL-FOXO4* (n=1), *PICALM-MLLT10* (n=1), *ETV6-NCOA2* (n=1) and *SET-NUP214* (n=1). In addition, we identified 3 novel chimeric transcripts in T-ALL, namely *NUP98-PSIP1* (n=1), *TPM3-JAK2* (n=1) and *SSBP2-FER* (n=1) and a novel *DDX3X-MLLT10* fusion transcript (n=1) recently described in a pediatric T-ALL patient [54].

Figure 5. *SSBP2-FER* and *TPM3-JAK2* fusions transform lymphoid cells and show constitutive activity (next page).
(**A**) Schematic representations of the predicted *SSBP2-FER* and *TPM3-JAK2* fusion joining the dimerization units of SSBP2 (LisH domain) or TPM3 (coiled-coil domains) to the TK domain of FER or JAK2, respectively. (**B**) Proliferation curve of mouse Ba/F3 cells in the absence of the cytokine interleukin 3 (IL3) (upper graph) and in the presence of ruxolitinib (lower graph). In the absence of IL3, cells expressing empty vector died whereas cells expressing the SSBP2-FER or TPM3-JAK2 fusion protein were transformed and could proliferate. Ba/F3 cells expressing the oncogenic JAK1 A634D mutant were used as positive control for transformation 18. The graph shows mean +/- st. dev. The lower graph illustrates the effects of the JAK kinase inhibitor ruxolitinib on Ba/F3 cell proliferation after 24 hours of treatment. The graph represents mean +/- st. dev. of triplicate measurements. (**C**) Western blot analysis of Ba/F3 cells transformed by the indicated kinases. The 2 upper panels show phosphorylation of the JAK and FER kinases, the panels below illustrate phosphorylation of downstream targets STAT5, STAT3, SRC and ERK1/2. (**D**) TCR gene fusions result in overexpression of a flanking gene in *RIC3-TRBC2* and *SFTA3-TRDC* fusions. The barplot is drawn for relative (to Thymus) expression values for the upstream and downstream flanking genes around *RIC3* and *SFTA3* for R4 and TLE90 samples, respectively. In both cases, the nearest downstream neighbor shows increased expression. (**E**) The heatmap illustrates the expression patterns of *RIC3* and *SFTA3*, together with their immediately upstream and downstream flanking genes in the genome, showing strong over-expression (red) of *LMO1* near the *RIC3* fusion, and of *NKX2-1* near the *SFTA3* fusion.


A
SSBP2
FER
SSBP2-FER
TPM3
JAK2
TPM3-JAK2
LisH
Pro rich
FCH
SH2
TK
Coiled coil
FERM
SH2
TK1
TK2
B
TPM3-JAK2
SSBP2-FER
JAK1 A634D
empty vector
Viable cells (x10⁶)
Days after cytokine removal
Relative proliferation (%)
Ruxolitinib (Mμ)
C
TPM3-JAK2
SSBP2-FER
JAK1 A634D
Phospho JAK1
Phospho TPM3/JAK1
Phospho SSBP2/FER
Phospho STAT5
Phospho STAT3
Phospho SRC
Phospho ERK1/2
D
log2(R4/Thymus)
CYB5R2 OVCH2 OR10AB1P OR5P2 OR5P3 OR10A6 OR10A3 NLRP10 EIF3F TUB RIC3 LMO1 STK33 TRIM66 RPL27A ST5 AKIP1 C11orf16 TMEM9B NRIP3
Upstream flanking genes
Downstream flanking genes
log2(TLE90/Thymus)
FAM177A1 PPP2R3C RP11-173D9.3 PSMA6 NFKBIA INSM2 RALGAPA1 BRMS1L MBIP SFTA3 NKX2-1 NKX2-8 PAX9 SLC25A21 RP11-537P22.1 MIPOL1 FOXA1 TTC6 SSTR1 CLEC14A
Upstream flanking genes
Downstream flanking genes
E
normalized log2(counts)
-2
6
R4 R5 R3 R1 R6 ALLSIL CCRFCEM HPBALL HSB2 JURKAT KARPAS45 KE37 LOUCY MOLT14 MOLT4 P12ICHIKAWA PEER RPMI8402 SUPT1 SUPT13 TALL1 R8 R9 R7 R10 TLE76 TLE77 TLE79 TLE80 TLE81 TLE85 TLE86 TLE87 TLE89 TLE90 TLE91 TLE92 TLE93 THYMUS DND41 TUG1 TUG2 TUG3 TUG4 TUG5 TUG6 TUG7 TUG8 RM2 PF382
TUB
RIC3
LMO1
MBIP
SFTA3
NKX2-1

Conventional cytogenetic analysis confirmed the presence of a t(X;10) in the case with the *DDX3X-MLLT10* fusion, whereas it failed to detect the chromosomal rearrangements for the *TPM3-JAK2, NUP98-PSIP1* and *SSBP2-FER* fusions, demonstrating the power of RNA-seq to identify cryptic fusion genes and to provide genetic information even in patients with uninformative cytogenetics. Reassuringly, RT-PCR and Sanger sequencing confirmed the presence of these fusion transcripts (**Table S12**).
The *TPM3-JAK2* and *SSBP2-FER* fusions encode typical tyrosine-kinase fusions that join the tyrosine-kinase domain of *JAK2* or *FER* to the dimerization units of *TPM3* or *SSBP2*, respectively (**Figure 5.A**). To assess whether the *TPM3-JAK2* and *SSBP2-FER* fusions encode oncogenic proteins, we tested their transforming properties in the *IL-3*–dependent Ba/F3 cell line [55]. Both *TPM3-JAK2* and *SSBP2-FER* transformed Ba/F3 cells to *IL-3*–independent growth, with even faster kinetics than the *JAK1* A634D mutant, which is a known transforming kinase [18] (**Figure 5.B**). Western blot analysis confirmed the constitutive autophosphorylation of the *JAK2* and *FER* fusion proteins, as well as the downstream STAT proteins (**Figure 5.C**). Ba/F3 cells transformed by the *TPM3-JAK2* fusion were sensitive to a JAK kinase inhibitor, documenting the potential application of *JAK2* kinase inhibitors for the treatment of T-ALL cases with *JAK2* fusion genes. No specific *FER* inhibitors were available to test their activity. Both *TPM3-JAK2* and *SSBP2-FER* fusion were screened in 50 additional T-ALL samples, but no additional case with these fusions was found.

Third, we also analyzed the identified fusions that did not seem to encode chimeric proteins (out-of-frame fusions), and which were the majority of fusions detected in T-ALL. These fusion events can be used as surrogate markers for the identification of chromosomal rearrangements, providing accurate information on the precise chromosomal breakpoints. In combination with the gene expression data obtained by RNA-seq, these data can identify genes that are located close to such potential breakpoints and for which the expression is significantly up- or downregulated. As expected, we identified the *STIL-TAL1* fusion in several T-ALL cases (n=8). We also identified and validated 6 fusion events involving TCR genes. In 4 of these cases, the TCR gene was found to be fused to the potential oncogene (*NOTCH1*, *IL7R*, *PLAG1*, *TLX1*). In the two other cases (R4, TLE90), the TCR gene was fused to *RIC3* or *SFTA3*, resulting in the ectopic expression of *LMO1* and *NKX2-1*, respectively, as indicated by RNA-seq gene expression data **(Figure 5.D and E**). Similarly, we could better characterize the t(10;14) in ALL-SIL cell line that expresses *TLX1* at high level.

In addition to the TCR gene rearrangements, also other fusions were associated with overexpression. We detected out-of-frame fusion transcripts that joined exon 4 of *CDK6* to exon 2 of *HOXA11-AS* and exon 5 of *CDK6* to sequences downstream of *EVX1*. In the same patient we also detected a fusion joining *DPY19L1* on

chromosome 7p14 to *HOXA11* on chromosome 7p15. The gene expression analysis documented high expression of genes of the *HOXA* cluster (i.e. *HOXA9, -A5, -A13, -A10, -A11*). Moreover, other fusions identified in this study, such as *CLINT1-MEF2C*, *HNRP-ZNF219* (n=2), *ZEB1-BMI1* and *AHI1-MYB* (n=2) were also associated with transcriptional activation of *MEF2C, ZNF219, BMI1* and *MYB* as confirmed by the expression data (**Table S9 and S12 and Figure S10**). Increased *MYB* expression in T-ALL was previously observed as a consequence of *MYB* duplication (including in the BE-13 cell line), which may also explain the detected *AHI1-MYB* fusion [8,56].

Finally, we also found out-of-frame fusion transcripts leading to the potential inactivation of tumor suppressor genes, such as *TP53-TBC1D3F* (ALLSIL cell line), *PTEN-RNLS* (LOUCY cell line), *IKZF1-ABCA13* and *CDKN2A-miR31HG* (R6 case), indicating a third class of fusion events (**Figure S10**). FISH analysis performed in the R6 case confirmed the p15/p16 deletion. As the genes are in close proximity, the *IKZF1-ABCA13* was presumably generated by deletion although no material was available to confirm this hypothesis.

DISCUSSION

The landscape of genomic variation underlying T-ALL has recently been investigated by sequencing candidate genes [14,21], whole exomes [17] and whole genomes [13]. The results of these studies, combined with a large body of gene-by-gene evidence collected over the last decade, provide a growing comprehension of the T-ALL genome. The T-ALL genome is mainly characterized by the over-expression of TF, such as *TLX1/3* and *TAL1*, in combination with gain-of-function *NOTCH1* mutations, and with additional mutations in chromatin modifiers, cellular signaling factors such as those involved in the JAK-STAT signaling pathway [57], tumor suppressor genes (*TP53, PTEN, WT1*), or in other genes such as ribosomal genes [17]. Since the majority of observed mutations are point mutations and gene fusions (much more than copy number aberrations [13]) we reasoned that RNA-seq would be effective to identify many of these mutations, certainly those associated with (over-)expressed oncogenes. Indeed, exome sequencing allows identifying point mutations but not gene fusions; and low coverage whole-genome sequencing allows identifying structural variation (gene fusions) but not point mutations. In this study we present RNA-seq analyses on a heterogeneous group of 31 T-ALL samples and 18 T-ALL cell-lines and demonstrate that RNA-Seq is indeed a very powerful approach to detect gene mutations and fusions as well as expression perturbations.

Our first challenge with regards to the accurate identification of point mutations was finding the optimal analysis pipeline – from read mapping to SNV calling and filtering – to avoid too many false positive SNVs. By exploiting whole-exome sequencing data for a subset of our samples we obtained a recovery ratio of 32%

when compared to the exome derived SNVs; a ratio that is comparable with previous RNA-seq studies [30,31]. However, this concordance could only be achieved by using the optimal read mapping methods and parameters: (1) use of a recent version of TopHat2 (v. 2.0.5. or higher) and (2) forcing this aligner to map all reads twice to the genome (once directly and once using split reads) and once to the transcriptome. Indeed, the computational task of sequence read mapping is more challenging for RNA-seq data because a large fraction of the obtained reads need to be split to allow reads that overlap exon-exon boundaries in the cDNA to be mapped to the genome. In this way, RNA-seq is more prone to the identification of false SNVs due to the erroneous mapping of reads, for example to highly similar non-spliced pseudogenes. For example, in the RPMI8402 cell line, 603 RNA-seq exclusive SNVs were found with the genome mapping strategy, while only 35 when using combined mapping strategy.

Among the previously published large scale RNA-seq cancer studies, only a handful performed variant calling on the RNA-seq data [30,31,58,59]. A combined mapping strategy was followed in all cases either by mapping the reads to a customized genome reference file (by the addition of exon junction segments) or mapping the reads twice (once to the genome and once to the transcriptome). Variant calling pipelines also showed diversity: Morin *et al* and Shah *et al* used SNVMix [60] for variant calling, while Seo *et al* and Berger *et al* implemented filters based on alignment on the non-reference bases. To our knowledge there is no extensive benchmarking study evaluating aligners and variant callers for RNA-seq data, but a review paper by Quinn *et al* compared the performance of two variant callers (GATK [23] and SAMTools[27]) with the optional duplicate removal step (pre and post alignment), and concluded that post-alignment duplicate removal and variant calling with SAMTools achieved the best performance in terms of sensitivity and specificity [61]. We have also followed the same strategy in our study and we could achieve a comparable recovery ratio of 32% when compared to Exome-seq calls.

A second challenge in identifying point mutations was the prioritization of candidate driver mutations versus passenger mutations. Due to the lack of matched germline RNA for each patient as control, we used a large cohort of local normal exome datasets, in combination with the commonly used variants from dbSNP and 1000genomes, to distinguish SNPs from candidate somatic mutations. This strategy has been successfully used before on transcriptome sequencing studies [62]. Identifying candidate cancer genes by gene mutation frequency is a frequently used approach [13,30,58]. Remarkably, by simply selecting all genes having a candidate somatic mutation in at least two samples (213 genes in total), we already achieved a highly significant enrichment for T-ALL related genes, such as *NOTCH1*, *BCL11B*, *FBXW7*, *DNM2*, *JAK3*, *JAK1*, and *IL7R*. Among the remaining candidates we searched for additional evidence and we propose seven additional candidate drivers because they are either "functionally similar" to the previously known drivers, or because they were mutated somatically at least once in another T-ALL cohort [17], or both. Six of these genes, namely *CIC*, *H3F3A, PTK2B, STAT5B, ANKRD1* and

HADHA have already been implicated in other cancers [63-70] while *DOCK2* has no association with cancer yet.

We found a remarkable clustering of molecular functions among the identified T-ALL driver genes, with enrichment for functions related to the regulation of gene expression. TFs and their co-factors play a central role in transcriptional regulation and these proteins are often mutated in T-ALL. Also, many of these play important roles in the normal T-cell developmental gene regulatory network [71], such as *NOTCH1, TLX1, TLX3, TAL1, BCL11B, CTCF, FOXO4, MYB*, and others. Upstream of these activated TFs, multiple kinases and other signaling factors control their activity, and these regulators are also often mutated in T-ALL (for example, *JAK1, JAK3*, and *IL7R*). Finally, chromatin modifiers and methylation factors are recurrently mutated and these can have both generally pervasive but also specific effects on the expression of oncogenes, such as *MYC* [72]. When multiple driver mutations are serially acquired, their combined effect will result in oncogenic expression profiles, whereby genes supporting a growth advantage increase and genes negatively affecting growth advantage (e.g., apoptosis, senescense) decrease in expression. It will be an interesting future challenge to draw the connections between the observed DNA mutations, the oncogenic program, and the final gene expression changes that we and others observe in T-ALL samples. Finally, it is likely that non-coding mutations, such as those in promoters, enhancers, microRNAs, and lncRNAs, add to the cancer-related gene regulatory network changes underlying leukemogenesis.

As mentioned above, only mutations in genes that are actively transcribed are detected, and this likely adds to the specificity of driver gene detection. On the other hand, this could also present a limitation of RNA-seq, because loss-of-function mutations in tumor suppressor genes may lead to nonsense-mediated decay, and as consequence low sequence coverage to call mutations. Based on our data however, this is not the case because we could detect *PHF6* mutations in up to 4/31 patient cases (13%), where exome sequencing identified *PHF6* mutations in 9/67 cases (13%) [17] and Zhang et al identified *PHF6* mutations in 24/106 cases by means of whole genome sequencing and capillary sequencing [13].
Interestingly, the gene expression information used above (i.e., read coverage to identify point mutations) can be further exploited at the quantitative level, similar to gene expression studies performed with microarray technology over the last 15 years. As many leukemia driver genes are characterized by changes in gene expression, this level of information is invaluable, both in research and diagnostic settings. We investigated how accurate gene expression levels can be achieved and we found that multiple normalization steps are required, both within-sample (gene length and gene GC content) and across samples (library size), and that batch effects can be effectively removed using a previously published Generalized Linear Model (GLM) [73]. The gene expression levels of the known drivers (e.g., *TLX1/3,*

TAL1, NOTCH1) are highly representative as driving events and as subtype identifiers. However, to discover driver genes *de novo*, using only gene expression values, is to our opinion not feasible (data not shown). Alternatively, we attempted to select candidate drivers based on the expression similarity (i.e., co-expression across the cohort) with known drivers. This led to the identification of *PTK2B*, whose expression strongly correlated with *JAK3* and which is known to be implicated in JAK-STAT signaling. The next level of gene expression analysis would preferably be a network-level analysis [74], but this requires a larger sample cohort.

Another kind of information that can be extracted from RNA-seq data, besides point mutations and gene expression changes, are alternative transcript events (ATE) and gene fusions [75]. We found only few significant ATEs but could confirm two exon-skipping events in the known T-ALL oncogenes SUZ12 and LCK. More importantly, we identified (i) known and novel in-frame fusions encoding chimeric proteins, (ii) TCR gene arrangements resulting in over-expression of oncogenes, and (iii) fusions not involving TCR genes but also resulting in over-expression of oncogenic transcription factors. The most recurrent fusion event, observed in 8/31 samples, was the *STIL-TAL1* fusion resulting in the ectopic over-expression of the *TAL1* gene. We also identified novel gene fusions, including two in-frame fusions, *TPM3-JAK2* and *SSBP2-FER,* producing chimeric oncoproteins; and other fusions resulting in the ectopic expression of transcription factors such as *PLAG1, MEF2C, ZNF219,* and *BMI1*. The ectopic expression of these genes is associated with a fusion event and with changed expression, which can both be detected by RNA-seq, making this technology extremely powerful to accurately detect such oncogenic events. Each of these novel events appears to be rare in T-ALL, as we identified at most 2 cases of each fusion. However the evidence of transcriptional activation of the partner genes suggests that further studies are required to establish the recurrence of these lesions and their functional meaning. It is notable that the normal thymus sample also shows four fusion events. However, as these genes are located in close proximity to each other, they may represent unannotated isoforms in the human transcriptome. Despite RNA-seq has offered a deeper insight into the complexity of the transcriptome, several studies have highlighted that the catalogue of all expressed transcripts is still far from complete and it is increasing the number of novel splice junctions connecting novel exon, non-exon regions, or linking independent transcripts [76].

Today, high-quality catalogues of driver genes across cancer types are available, and this influences how and why cancer genomes need to be sequenced. For T-ALL, and for many common cancer types, the objectives of sequencing are shifting from the discovery of cancer genes, to a diagnostic setting in which a list of driver events are *a priori* known. Targeted re-sequencing provides an interesting route, although this poses technical challenges of amplification or capturing, and perhaps more importantly, is focused on a limited number of genes and on one particular

mutation type, namely point mutations and small indels. We have shown in this study that, with a list of interesting cancer drivers at hand, and with other datasets being available (e.g., rare variants from local exome studies, 1000 genomes, TCGA data, etc), RNA-sequencing of only the cancer sample provides a technically straightforward approach and delivers at once the point mutations, gene fusions and gene expression changes across the entire transcriptome. And as a corollary, the data analysis strategies provided here would be beneficial for any cancer type as long as a body of knowledge is available for selecting and prioritizing candidate events.

MATERIALS AND METHODS

Patient samples and cell lines

Diagnostic total RNAs from 31 T-ALL patients (20 adults and 11 children) were collected at various institutions. All patients have given their informed consent and all samples were obtained according to the guidelines of the local ethical committees. This study was approved by the ethical committee of the University Hospital Leuven. Diagnosis of T-ALL was based on morphology, cytochemistry and immunophenotyping according to the World Health Organization and European Group for the Immunological Characterization of Leukemia criteria [77]. The clinical and hematologic features of the 31 patients at the diagnosis are summarized in Table S11 Total RNAs from 18 T-ALL cell lines (DSMZ, Braunschweig, Germany) were extracted using QIAGEN RNeasy Mini Kit. A pool of total RNAs from 5 normal human thymuses was purchased from Capital Biosciences.
All the RNA samples showed a high quality RNA Integrative Number (RIN>/=7) score on the Bioanalyzer (Agilent Technologies).
Fifty additional RNA samples were used for *TPM3-JAK2* and *SSBP2-FER* analysis. Genomic DNA from of 71 adult T-ALL patients were used for *H3F3A K28* screening.

RNA-seq

Next generation sequencing libraries were constructed from 500 ng of total RNA using the Truseq RNA sample prep kit (Illumina). RNA-seq libraries were subjected to 2 x 100bp paired-end sequencing on a HiSeq2000 instrument (Illumina). Sequence reads were processed to identify gene fusion transcripts, single nucleotide variants (SNVs) and gene expression levels. For the read mapping, variant calling and transcriptome assembly, we used the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules foundation and the Flemish Government - department EWI.

Fusion transcript discovery

Fusion transcript discovery was performed using defuse v.0.5.0 [49] with default parameters. The resulting list was filtered as described in [78]. Briefly, fusion transcripts with less than 8 spanning reads and less than 5 split reads were filtered

out. In addition, we removed fusion events observed in adjacent genes and fusion events involving ribosomal genes (ribosomal genes were downloaded from Biomart on 24-05-2011 using GO:0005840) and the genes located on chrM. Fusion events were annotated using Pegasus (http://sourceforge.net/projects/pegasus-fus/).

Gene expression analysis

For Gene Expression Profiling analysis, reads were mapped to the human reference genome (assembly GRCh37.68) using TopHat v.2.0.5 [26] with the following parameters: transcriptome-only. Read counts per gene were obtained with the HTSeq package (htseq-count) (http://www-huber.embl.de/users/anders/HTSeq). The aggregated read counts were normalized with EDASeq v1.4.0 [79] and generalized linear model was fitted with edgeR v3.0.4 [73] to remove batch effect originating from the sample collection center. The pathways, and upstream regulators were generated through the use of IPA (Ingenuity® Systems, www.ingenuity.com). Expression neighbors were detected with Pavlidis Template Matching (PTM) analysis [80]. Transcript based gene expression values were obtained using Cufflinks suite [81,82]. Transcript assembly was performed with cufflinks v2.1.1 with –g option using assembly GRCh37.68.

Gene set enrichment analysis (GSEA) was performed for TAL1, TLX and LYL1 clusters [83]. We have obtained whole genome rankings for *TAL1*, TLX (*TLX1* and *TLX3*), and *LYL1* simply by calculating the log fold changes between samples expressing the respective gene versus the remaining samples. The gene signatures from Soulier *et al* were obtained from Table S2 [41].

Alternative Transcript Event Discovery

Tumor patient samples and Thymus RNA-Seq samples were mapped to the Ensembl GRCh37.68 reference genome by Tophat2 [26]. Mapped reads were realigned, and transcript abundance were estimated using cufflinks v2.1.1 [81,82]. Transcript assembly was reconstructed using the cuffmerge program of the cufflinks package from the realigned transfrags for each of patient RNA-seq samples, merged with the Thymus sample (control), followed by differential expression analysis performed using cuffdiff2 program. The significant events were extracted from the list of differentially expressed genes, isoforms, primary transcripts and coding sequence and assessed manually with IGV [84]. The mRNA sequences for novel *SUZ12* and *LCK* transcripts were extracted using *gffread* command of cufflinks, and these sequences were translated using the *translate* tool of the ExPASy Bioinformatics Resource Portal [85]. The longest ORF sequence was used to verify the domain architecture of the resulting proteins using SMART [86,87].

Prediction of Single Nucleotide Variation

The sequence reads were mapped to the human reference genome (assembly GRCh37.68) using TopHat2 setting the option "read-realign-edit-dist" to zero [26]. Duplicate removal process was performed on the aligned reads using Picard v1.74

(http://picard.sourceforge.net). Then SAMTools package v0.1.19+ (pulled from the git repository on 29-07-2013) [27] was used for single nucleotide variant (SNV) and small insertion and deletion (INDEL) detection with minimum mapping quality threshold of 1 and minimum base quality threshold of 13 (-q 1 -Q 13) [27]. The variant calling was done on the coding regions of the genome only (extracted from the transcript definitions in the assembly GRCh37.68). The variant predictions that were supported exclusively by variants located in the beginning or the end of the read were filtered out. Then the SNVs were further filtered with depth of coverage threshold of 20 and minimum variant allele frequency threshold of 0.20. INDELs predictions were filtered with the SAMTools recommended parameters (varFilter -10 -20 -30 -40 -a4 -G90 -S30) and additionally INDELs located in homopolymer stretches longer than 5 bps were filtered. The high quality list of variants was filtered for common population variants using the calls from 1000 genomes, dbSNP, HapMap, and Complete Genomics. Note that, the list of common population variants was cleaned from oncogenic variants using COSMIC listed variants (v66) [36]. Moreover, the variants located in the repeat regions (simple repeat and RepeatMasker) were filtered out. Finally, the variants that are observed in the exomes of remission (i.e. healthy) samples (including the previously published 39 exome remissions [17] and the 6 additional exome remission sequenced) and the variants that are observed in Thymus were also filtered out. The final filtered list of variants was annotated with the Variant Effect Predictor version 2.7 [25] and the protein-altering mutations were selected. The following terms were used for selecting protein-altering SNV: splice_donor_variant, splice_acceptor_variant, stop_gained, initiator_codon_variant, missense_variant, splice_region_variant. The same terms were used for filtering the INDELs with the addition of the following terms: inframe_insertion, inframe_deletion, frameshift_variant.

The list of candidate genes was created by intersecting the genes with recurrent mutations (SNVs and INDELs) in RNA-seq patient cohort with the somatic mutations in Exome-seq patient cohort [17]. The list of genes that have recurrent mutations in the RNA-seq patient cohort was filtered for mutations observed in chrM.

The list of T-ALL driver genes were curated using the Census database [46] and T-ALL literature and includes the following genes: *TLX1, TLX3, PHF6, MYC, BCL11B, HOXA1, SET, MLL, MLLT1, PICALM, MLLT10, WT1, MYB, LEF1, LMO2, LMO1, TAL1, NUP98, NOTCH1, FBXW7, CCND2, PTEN, PTPN2, NF1, FLT3, JAK1, NRAS, LCK, NUP214, ABL1, EZH2, SETD2, SUZ12, JAK3, MEF2C, NKX2-1, NKX2-2, CDKN2A, CDKN2B, RUNX1, KRAS, EED, ETV6, RPL10, DNM2, IL7R, CNOT3.*

Exome-seq analysis

Somatic mutations from the exome pairs were obtained as described previously [17]. Briefly, the alignment was performed with BWA [22] and post-alignment modifications (duplicate removal, realignment around indels and calibration of the quality scores)

were done with the Genome Analysis Toolkit (GATK) [23]. Variant calling was performed with GATK using Variant Quality Score Recalibration (VQSR) method. Putative somatic variants were identified by subtracting the mutations observed in the primary samples from the mutations observed in the corresponding remission samples. SomaticSniper score above 70 was used to identify the final list of somatic events [24].
Variant allele frequency (VAF) plots were drawn for the positions that are novel SNVs in either of the RNA-seq or Exome-seq data and covered by at least 20 reads in both datasets.

RT-PCR and Sanger Sequencing
Novel candidate fusion transcripts were validated by Reverse-Transcription Polymerase-Chain-Reaction (RT-PCR) and Sanger sequencing. In all cases thymus was used as negative control. cDNA synthesis and PCR amplification were performed using standard protocols that come with Superscript III Reverse Transcriptase (Invitrogen) and GoTaq (Promega). PCR primers were designed to amplify 200-400 bp fragments containing the fusion boundary detected by RNA-seq. The PCR products were analyzed using a QIAxcel automated multicapillary electrophoresis system (QIAGEN). The results were processed and visualized using the BioCalculator Software. PCR products were analyzed by Sanger Sequencing. In cases where multiple PCR products were detected, we performed conventional agarose gel electrophoresis and extraction of specific bands using the gel DNA Recovery Kit (Zymo). Analysis of Sanger chromatograms was performed using CLC Main Workbench 6 (CLC Bio, Aarhus, Denmark). Fusion detection was performed using NCBI Blast alignment. Analysis of the breakpoint was done on the longest isoform reported on the Ensembl genome browser. The tested fusions predictions and the primers used for validations are reported in Table S12.
Validation of *SUZ12* exon skipping was performed by RT-PCR, gel extraction and sequencing of the two PCR products (**Figure 4.I**). The following primers were used for RT-PCR and Sanger sequencing: SUZ12_EX1F (CTGACCACGAGCTTTTCCTC) and SUZ12_EX9R (CCATTTCCTGCATGGCTACT).

Cloning
The plasmid *TPM3-JAK2* pMSCV-GFP was obtained as follows: a DNA fragment containing *TPM3* coding region till exon 7 was PCR amplified from thymus cDNA using Phusion High Fidelity DNA Polymerase (Finzyme) and primers containing BglII and XhoI restriction sites. Primers containing XhoI and EcoRI restriction sites were used to amplify *JAK2* coding exons 17-25. PCR products were cloned into the BglII and EcoRI sites of the pMSCV-GFP vector after subcloning into the pJET1.2 CloneJET vector (Fermentas). As a final control, plasmid DNA was sequenced by Sanger sequencing.

SSBP2-FER fusion was synthesized by Genscript (Piscataway, NJ, USA) and cloned into pMSCV-GFP by using the unique restriction sites XhoI and EcoRI. The plasmid contained the full length *SSBP2-FER* fusion including the first 16 coding exons of *SSBP2* and the coding exons 14-20 of *FER*.

Cell culture
Viral supernatants were produced in HEK293T cells using an EcoPack packaging plasmid and TurboFect transfection reagent (Fermentas). Viruses were harvested 48 hours after transfection followed by transduction of the Ba/F3 murine pro-B cells (DSMZ, Braunschweig, Germany) as described previously [88].

Transformation experiments
Ba/F3 cells were washed twice in PBS to remove all traces of cytokines and were seeded in triplicate in 24-well dishes at 100 000 cells/mL. GFP expression and cell number were measure on a Guava flow cytometer (Millipore). All experiments were terminated at day 8 after cytokine removal and cell lines showing no sign of cell proliferation at that timepoint were declared to be non-transforming.

Western blotting
Total cell lysates were analyzed by standard electrophoresis and western blotting procedures using the following antibodies: anti-phospho-*JAK1* (Tyr1022/1023), anti-phospho-*STAT1*, anti–phospho-*STAT5* (Tyr694), anti–phospho-*STAT3* (Tyr705), anti-phospho*ERK1*-2, anti-phospho-SRC families (Tyr416) (from Cell Signaling Technology).

Inhibitor experiments
TPM3-JAK2 and *SSBP2-FER* IL3-independent Ba/F3 cells were seeded in triplicate in 96-well plates at a density of 0.03 X 10^6 cells in the presence of *JAK* inhibitor Ruxolitinib (INCB018424, Azon Medchem). Cell proliferation and viability were assessed on a Guava flow cytometer after 24 hours to determine the IC50, the concentration of inhibitor that gave a 50% inhibition.

ACCESSION NUMBERS
Genome data has been deposited at the European Genome-phenome Archive (EGA,http://www.ebi.ac.uk/ega/) which is hosted at the EBI, under accession number EGAS00001000536.

AUTHOR CONTRIBUTIONS
ZKA, GH, EG, KDK, NM, and VG performed experiments and analyzed the data; JC and SA conceived the study and analyzed the data; SC, IW, JCloos, RF, and FS provided materials. ZKA, VG, and SA wrote the manuscript. All authors contributed to the manuscript.

REFERENCES

1. Pieters, R. & Carroll, W. L. Biology and treatment of acute lymphoblastic leukemia. *Pediatr. Clin. North Am.* **55,** 1–20– ix (2008).
2. Van Vlierberghe, P. & Ferrando, A. The molecular basis of T cell acute lymphoblastic leukemia. *J. Clin. Invest.* **122,** 3398–3406 (2012).
3. Graux, C., Cools, J., Michaux, L., Vandenberghe, P. & Hagemeijer, A. Cytogenetics and molecular genetics of T-cell acute lymphoblastic leukemia: from thymocyte to lymphoblast. *Leukemia* **20,** 1496–1510 (2006).
4. Le Noir, S. *et al.* Extensive molecular mapping of TCRα/δ- and TCRβ-involved chromosomal translocations reveals distinct mechanisms of oncogene activation in T-ALL. *Blood* **120,** 3298–3309 (2012).
5. Van Vlierberghe, P. *et al.* Cooperative genetic defects in TLX3 rearranged pediatric T-ALL. *Leukemia* **22,** 762–770 (2008).
6. Ferrando, A. A. *et al.* Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* **1,** 75–87 (2002).
7. Sulong, S. *et al.* A comprehensive analysis of the CDKN2A gene in childhood acute lymphoblastic leukemia reveals genomic deletion, copy number neutral loss of heterozygosity, and association with specific cytogenetic subgroups. *Blood* **113,** 100–107 (2009).
8. Lahortiga, I. *et al.* Duplication of the MYB oncogene in T cell acute lymphoblastic leukemia. *Nat Genet* **39,** 593–595 (2007).
9. Graux, C. *et al.* Fusion of NUP214 to ABL1 on amplified episomes in T-cell acute lymphoblastic leukemia. *Nat Genet* **36,** 1084–1089 (2004).
10. Weng, A. P. *et al.* Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* **306,** 269–271 (2004).
11. Shochat, C. *et al.* Gain-of-function mutations in interleukin-7 receptor-α (IL7R) in childhood acute lymphoblastic leukemias. *Journal of Experimental Medicine* **208,** 901–908 (2011).
12. Zenatti, P. P. *et al.* Oncogenic IL7R gain-of-function mutations in childhood T-cell acute lymphoblastic leukemia. *Nat Genet* **43,** 932–939 (2011).
13. Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481,** 157–163 (2012).
14. Kalender Atak, Z. *et al.* High accuracy mutation detection in leukemia on a selected panel of cancer genes. *PLoS ONE* **7,** e38463 (2012).
15. Bains, T. *et al.* Newly described activating JAK3 mutations in T-cell acute lymphoblastic leukemia. *Leukemia* **26,** 2144–2146 (2012).
16. Elliott, N. E. *et al.* FERM domain mutations induce gain of function in JAK3 in adult T-cell leukemia/lymphoma. *Blood* **118,** 3911–3921 (2011).
17. De Keersmaecker, K. *et al.* Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nat Genet* **45,** 186–190 (2013).
18. Flex, E. *et al.* Somatically acquired JAK1 mutations in adult acute

lymphoblastic leukemia. *Journal of Experimental Medicine* **205,** 751–758 (2008).

19. Porcu, M. *et al.* Mutation of the receptor tyrosine phosphatase PTPRC (CD45) in T-cell acute lymphoblastic leukemia. *Blood* **119,** 4476–4479 (2012).
20. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11,** 685–696 (2010).
21. Van Vlierberghe, P. *et al.* PHF6 mutations in T-cell acute lymphoblastic leukemia. *Nat Genet* **42,** 338–342 (2010).
22. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
23. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491–498 (2011).
24. Larson, D. E. *et al.* SomaticSniper: Identification of Somatic Point Mutations in Whole Genome Sequencing Data. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btr665
25. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).
26. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14,** R36 (2013).
27. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
28. Bass, B. *et al.* The difficult calls in RNA editing. Interviewed by H Craig Mak. *Nature Biotechnology* **30,** 1207–1209 (2012).
29. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods* **8,** 469–477 (2011).
30. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486,** 395–399 (2012).
31. Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476,** 298–303 (2011).
32. Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Res* **12,** 656–664 (2002).
33. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29,** 308–311 (2001).
34. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
35. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2010).
36. Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10,** Unit 10.11 (2008).
37. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).

38. Sturm, D. *et al.* Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* **22,** 425–437 (2012).
39. Aerts, S. *et al.* Gene prioritization through genomic data fusion. *Nature Biotechnology* **24,** 537–544 (2006).
40. Bash, R. O. *et al.* Does activation of the TAL1 gene occur in a majority of patients with T-cell acute lymphoblastic leukemia? A pediatric oncology group study. *Blood* **86,** 666–676 (1995).
41. Soulier, J. *et al.* HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood* **106,** 274–286 (2005).
42. Miyazaki, T. *et al.* Pyk2 is a downstream mediator of the IL-2 receptor-coupled Jak signaling pathway. *Genes Dev.* **12,** 770–775 (1998).
43. Gardina, P. J. *et al.* Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics* **7,** 325 (2006).
44. Thorsen, K. *et al.* Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. *Mol. Cell Proteomics* **7,** 1214–1224 (2008).
45. Guttery, D. S., Shaw, J. A., Lloyd, K., Pringle, J. H. & Walker, R. A. Expression of tenascin-C and its isoforms in the breast. *Cancer Metastasis Rev.* **29,** 595–606 (2010).
46. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4,** 177–183 (2004).
47. Cauwelier, B. *et al.* Molecular cytogenetic study of 126 unselected T-ALL cases reveals high incidence of TCRbeta locus rearrangements and putative new T-cell oncogenes. *Leukemia* **20,** 1238–1244 (2006).
48. Oram, S. H. *et al.* Bivalent promoter marks and a latent enhancer may prime the leukaemia oncogene LMO1 for ectopic expression in T-cell leukaemia. *Leukemia* (2013). doi:10.1038/leu.2013.2
49. McPherson, A. *et al.* deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol* **7,** e1001138 (2011).
50. Nacu, S. *et al.* Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med Genomics* **4,** 11 (2011).
51. Zhou, J., Liao, J., Zheng, X. & Shen, H. Chimeric RNAs as potential biomarkers for tumor diagnosis. *BMB Rep* **45,** 133–140 (2012).
52. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37,** 1–13 (2009).
53. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4,** 44–57 (2009).
54. Brandimarte, L. *et al.* New MLLT10 gene recombinations in pediatric T-acute lymphoblastic leukemia. *Blood* (2013). doi:10.1182/blood-2013-02-487256

55. Warmuth, M., Kim, S., Gu, X.-J., Xia, G. & Adrián, F. Ba/F3 cells and their use in kinase drug discovery. *Curr Opin Oncol* **19,** 55–60 (2007).
56. O'Neil, J. *et al.* Alu elements mediate MYB gene tandem duplication in human T-ALL. *Journal of Experimental Medicine* **204,** 3059–3066 (2007).
57. Vainchenker, W. & Constantinescu, S. N. JAK/STAT signaling in hematological malignancies. *Oncogene* **32,** 2601–2613 (2013).
58. Seo, J.-S. *et al.* The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res* **22,** 2109–2119 (2012).
59. Berger, M. F. *et al.* Integrative analysis of the melanoma transcriptome. *Genome Res* **20,** 413–427 (2010).
60. Goya, R. *et al.* SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26,** 730–736 (2010).
61. Quinn, E. M. *et al.* Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. *PLoS ONE* **8,** e58815 (2013).
62. Liu, J. *et al.* Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events. *Genome Res* **22,** 2315–2327 (2012).
63. Bettegowda, C., Agrawal, N., Jiao, Y., Sausen, M. & Wood, L. D. Mutations in CIC and FUBP1 contribute to human oligodendroglioma. *Science* (2011).
64. Schwartzentruber, J. *et al.* Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature* **482,** 226–231 (2012).
65. Wu, G. *et al.* Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. *Nat Genet* **44,** 251–253 (2012).
66. Sun, C. K. *et al.* Proline-rich tyrosine kinase 2 (Pyk2) promotes proliferation and invasiveness of hepatocellular carcinoma cells through c-Src/ERK activation. *Carcinogenesis* **29,** 2096–2105 (2008).
67. Sun, C. K. *et al.* Proline-rich tyrosine kinase 2 (Pyk2) promotes cell motility of hepatocellular carcinoma through induction of epithelial to mesenchymal transition. *PLoS ONE* **6,** e18878 (2011).
68. Rajala, H. L. M. *et al.* Discovery of somatic STAT5b mutations in large granular lymphocytic leukemia. *Blood* (2013). doi:10.1182/blood-2012-12-474577
69. Noll, J. E. *et al.* Mutant p53 drives multinucleation and invasion through a process that is suppressed by ANKRD11. *Oncogene* **31,** 2836–2848 (2012).
70. Mamtani, M. & Kulkarni, H. Association of HADHA expression with the risk of breast cancer: targeted subset analysis and meta-analysis of microarray data. *BMC Res Notes* **5,** 25 (2012).
71. Kueh, H. Y. & Rothenberg, E. V. Regulatory gene network circuits underlying T cell development from multipotent progenitors. *Wiley Interdiscip Rev Syst Biol Med* **4,** 79–102 (2012).
72. Uribesalgo, I., Benitah, S. A. & Di Croce, L. From oncogene to tumor suppressor: the dual role of Myc in leukemia. *Cell Cycle* (2012).

73. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
74. Carro, M. S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463,** 318–325 (2010).
75. Maher, C. A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458,** 97–101 (2009).
76. Halvardson, J., Zaghlool, A. & Feuk, L. Exome RNA sequencing reveals rare and novel alternative transcripts. *Nucleic Acids Res* **41,** e6 (2013).
77. Bene, M. C. *et al.* Proposals for the immunological classification of acute leukemias. European Group for the Immunological Characterization of Leukemias (EGIL). in *Leukemia* **9,** 1783–1786 (1995).
78. Steidl, C. *et al.* MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. *Nature* **471,** 377–381 (2011).
79. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12,** 480 (2011).
80. Gillis, J. & Pavlidis, P. Assessing identity, redundancy and confounds in Gene Ontology annotations over time. *Bioinformatics* **29,** 476–482 (2013).
81. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28,** 511–515 (2010).
82. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27,** 2325–2329 (2011).
83. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 15545–15550 (2005).
84. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29,** 24–26 (2011).
85. Artimo, P. *et al.* ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res* **40,** W597–603 (2012).
86. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U.S.A.* **95,** 5857–5864 (1998).
87. Letunic, I., Doerks, T. & Bork, P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* **40,** D302–5 (2012).
88. De Keersmaecker, K. *et al.* Fusion of EML1 to ABL1 in T-cell acute lymphoblastic leukemia with cryptic t(9;14)(q34;q32). *Blood* **105,** 4849–4852 (2005).

SUPPLEMENTARY MATERIAL

Figure S1. Suboptimal mapping strategies result in incorrect read alignment.
Alignment of the Exome-seq and RNA-seq reads on *GLUD2* and *GLUD1* genes for the RPMI8402 cell line. Two alignment strategies are visualized in these figures for RNA-seq: genome-only mapping and combined mapping strategy. Panel (**A**) shows the alignment for *GLUD2* gene. With exome-seq a very high coverage was achieved (the coverage track scale is 0-1000). Aligning the RNA-seq reads with 'genome-only' option yields high coverage as well however with a lot of mismatches in the alignment (colored lines indicate the presence of a nucleotide different than the reference base). However, when combined mapping strategy is applied the coverage drops drastically. Panel (**B**) shows the alignment of *GLUD1* gene. When mapping with genome only option, the coverage is not high (the coverage track scale is 0-900) since the reads are forced to map to the pseudogene (*GLUD2*) with a lot of mismatched. When the combined mapping strategy implemented, the reads align to *GLUD1* gene correctly with less mismatches.

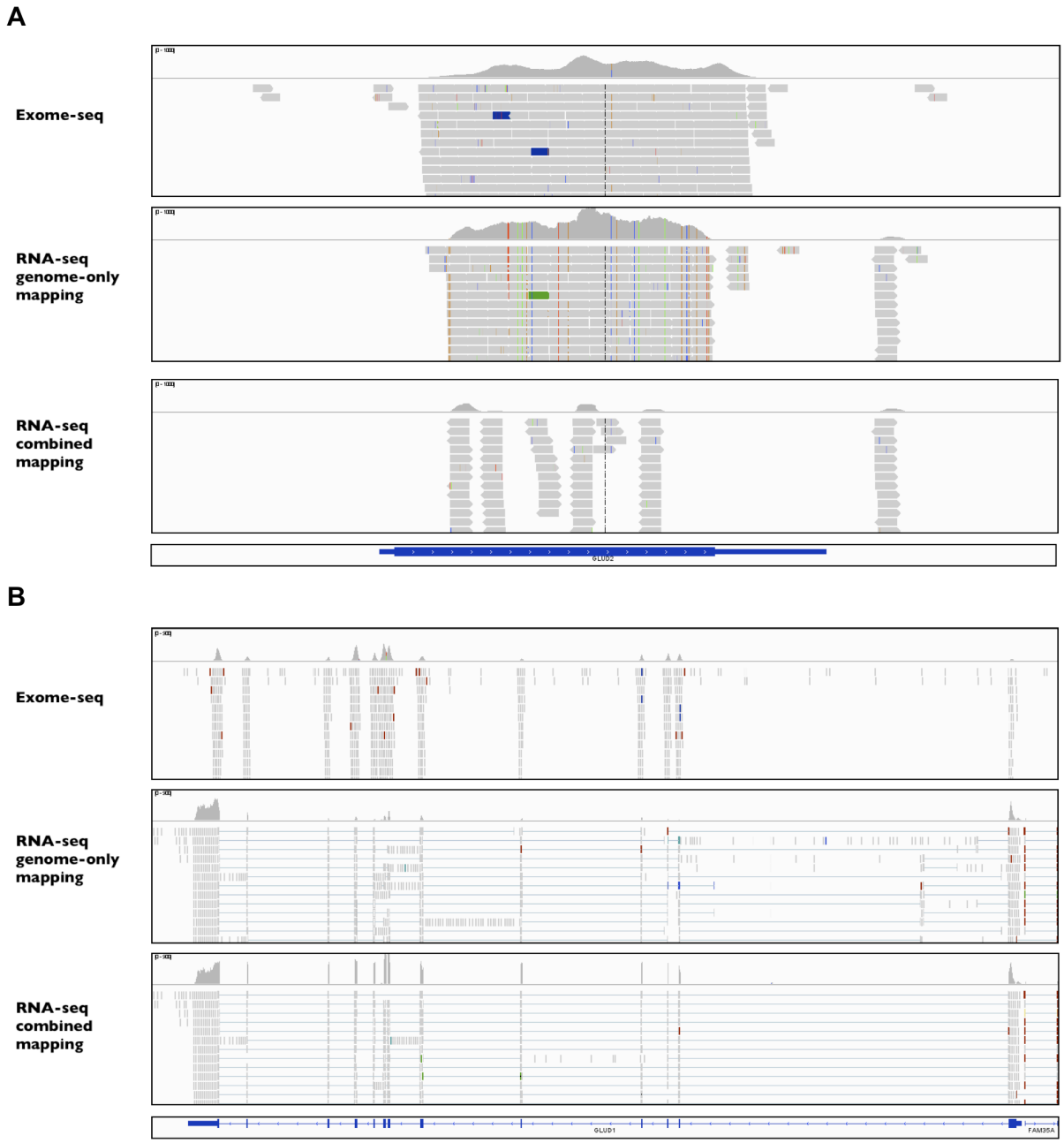


Figure S2. Variant allele frequency plots for assessing transcriptome-only mapping strategy.
The variant allele frequencines of the SNVs that have at least 20X reads in exome-seq and RNA-seq are plotted. The RNA-seq SNVs were obtained with the transcriptome-only alignment option. Red and green dots represent the SNVs that are detected only in RNA-seq and only in exome-seq, respectively, while black dots represent the SNVs that are called in both. Venn diagrams are produced from the points represented in the graphs. The plots are generated for (**A**) RPMI8402 cell line and (**B**) TLE79 patient sample.

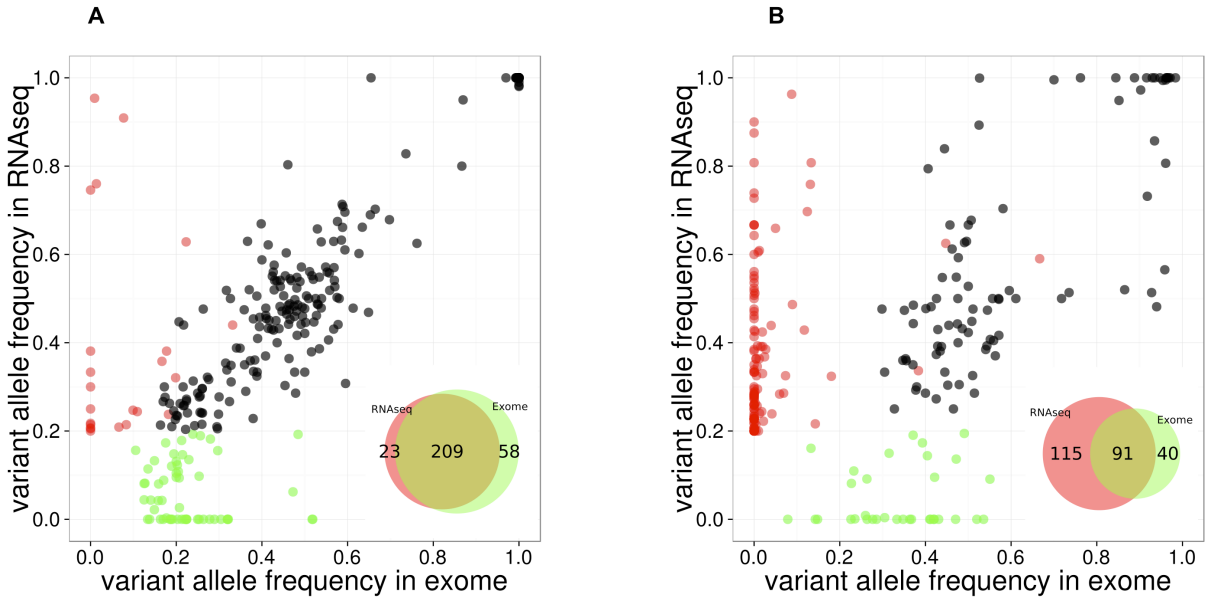
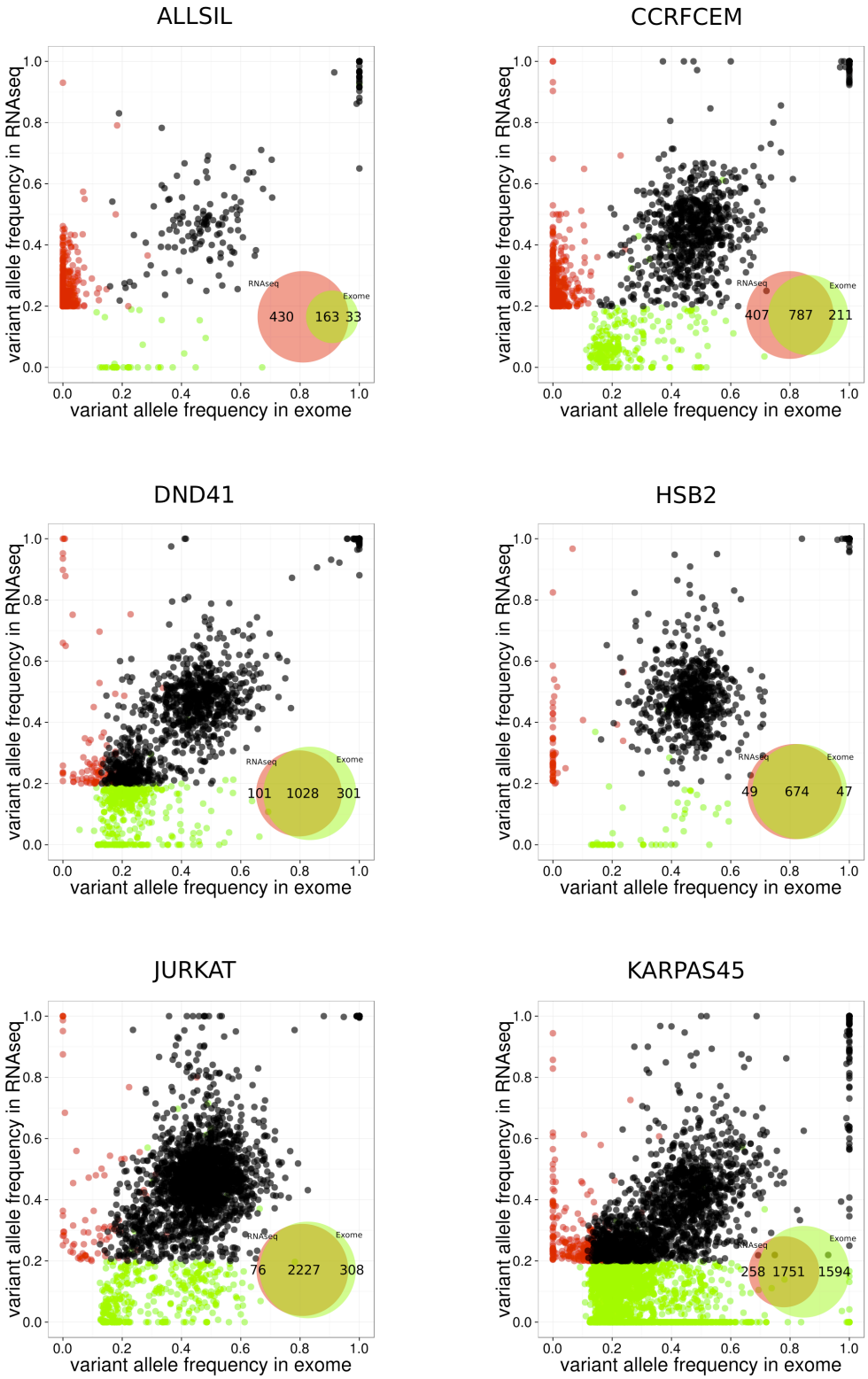
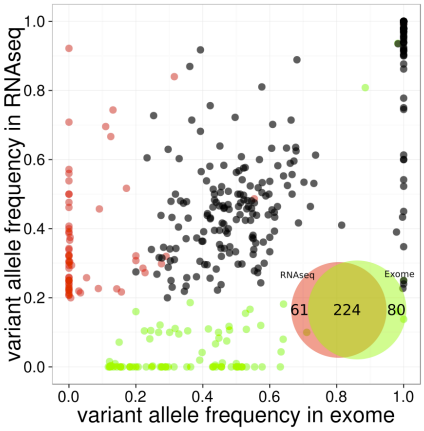


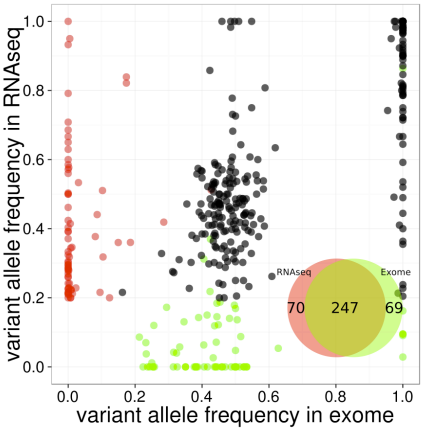
Figure S3. Variant Allele Frequency (VAF) plots for 16 cell lines and 20 patient samples. RNA-seq calls are made with combined mapping strategy. The venn diagrams and VAF plots are drawn for variants that have sequence coverage of at least 20X.



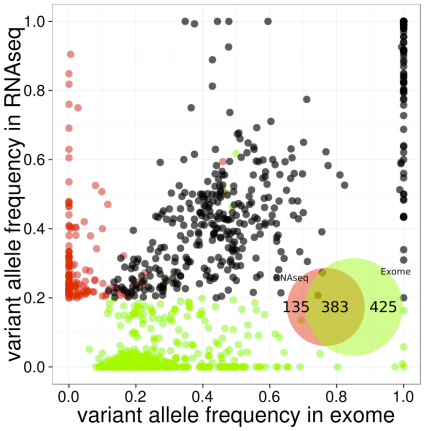
KE37



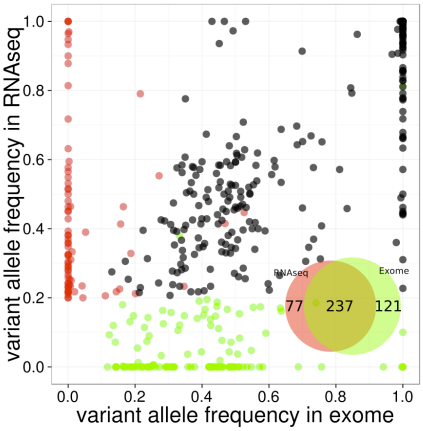
LOUCY



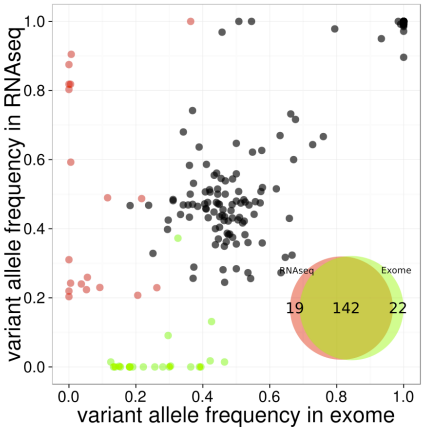
MOLT4



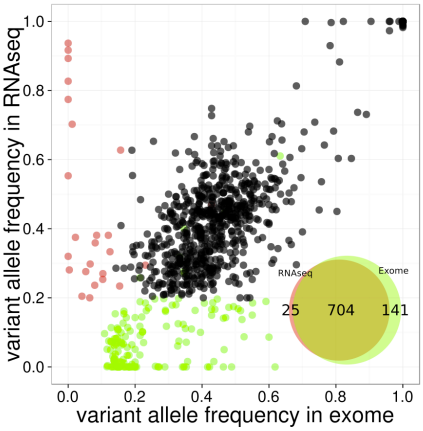
P12ICHIKAWA



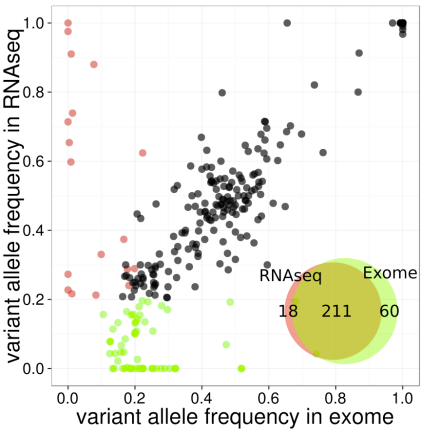
PEER



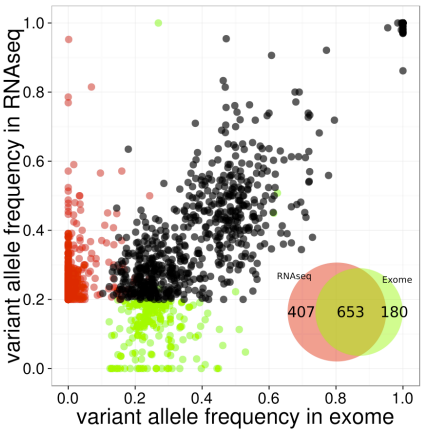
PF382



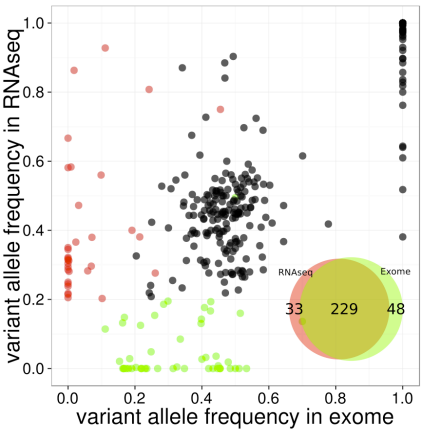
RPMI8402



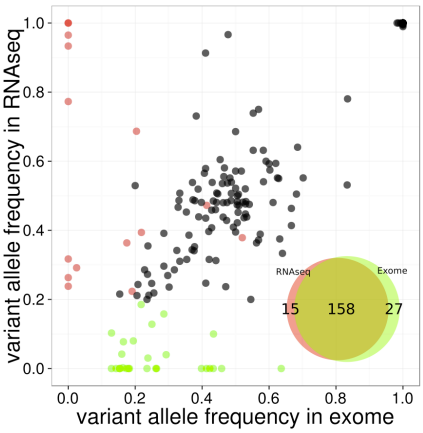
SUPT1



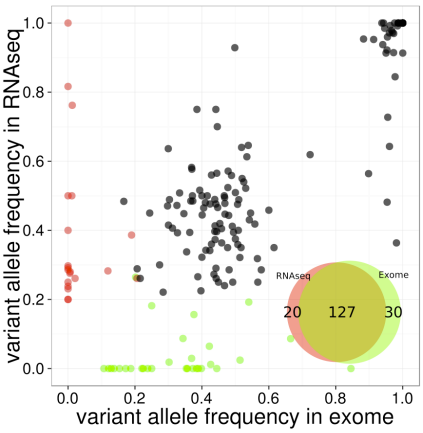
SUPT13



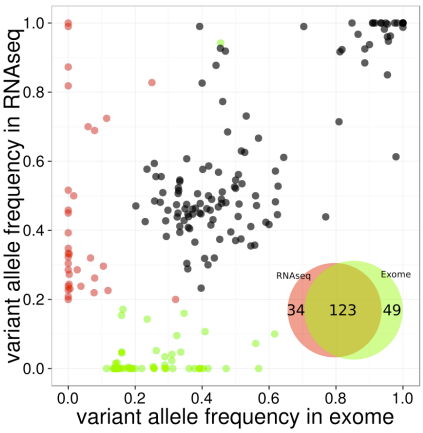
TALL1



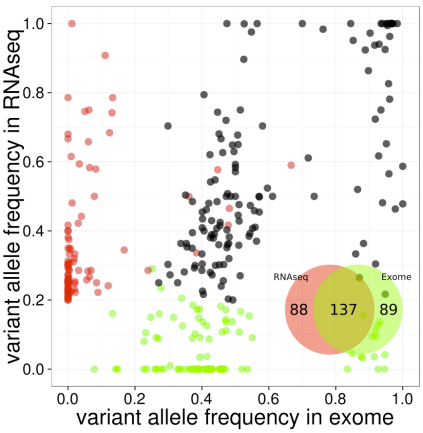
TLE76



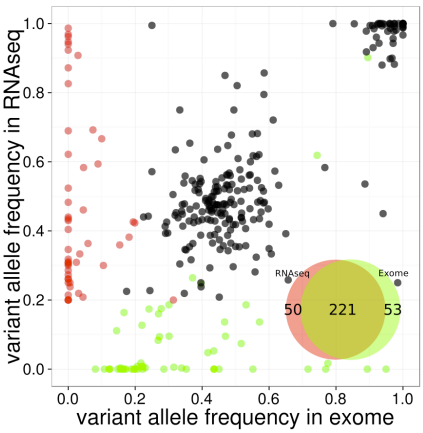
TLE77



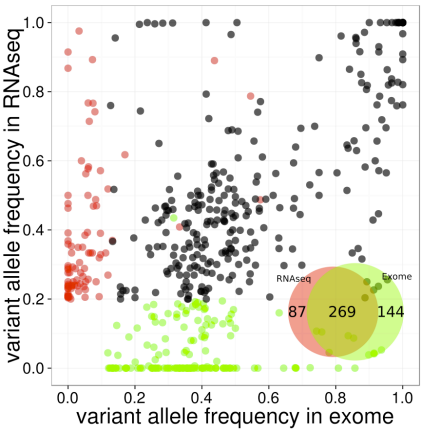
TLE79



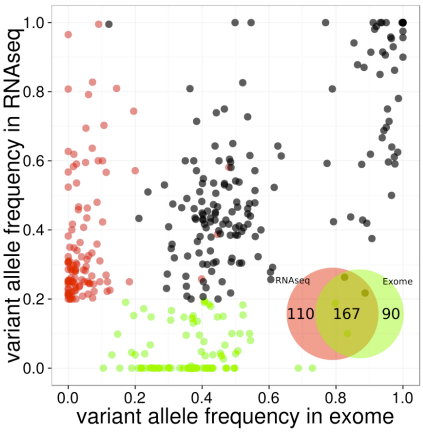
TLE80



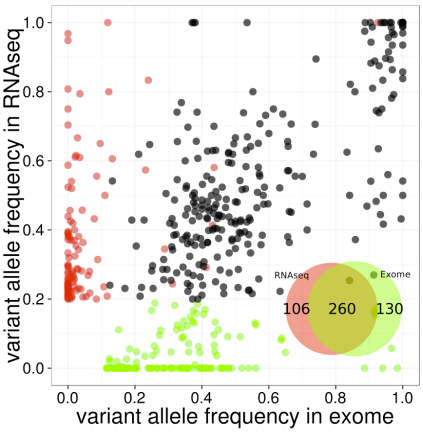
TLE81



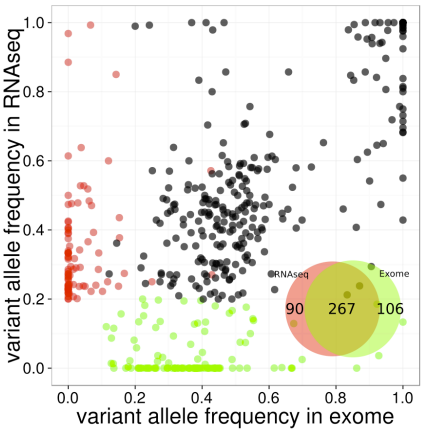
TLE85



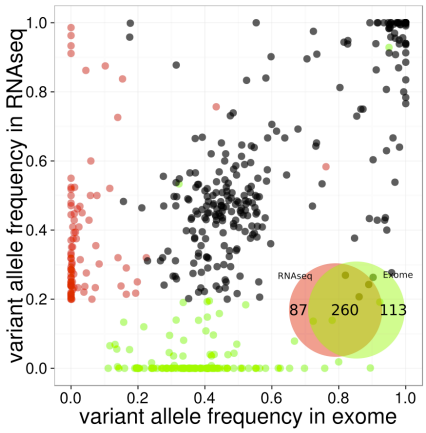
TLE86



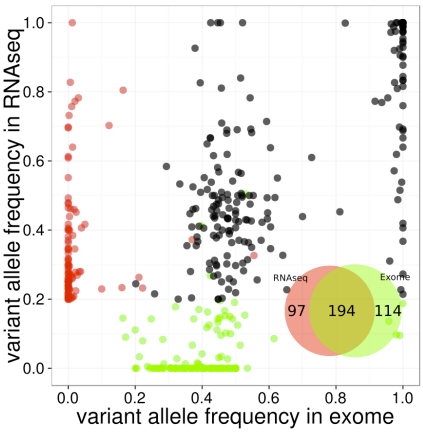
TLE87



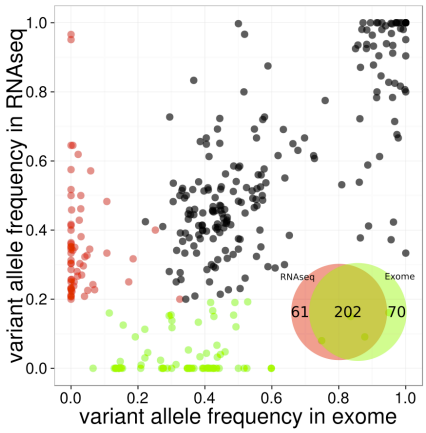
TLE89



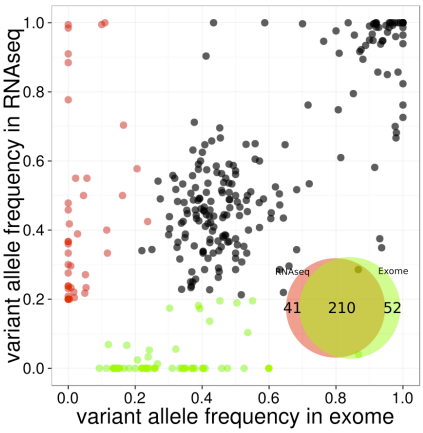
TLE90



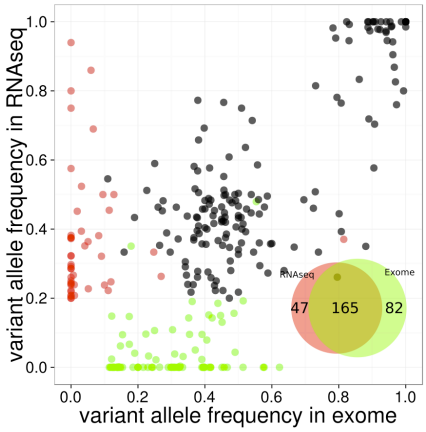
TLE91



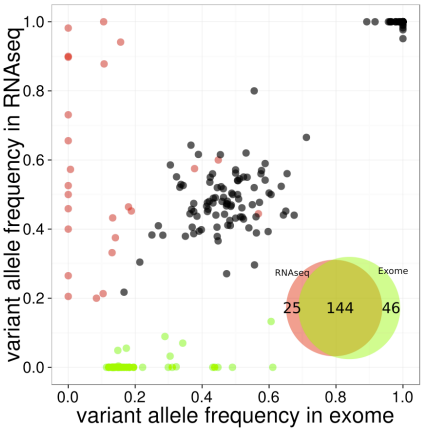
TLE92



TLE93



TUG1



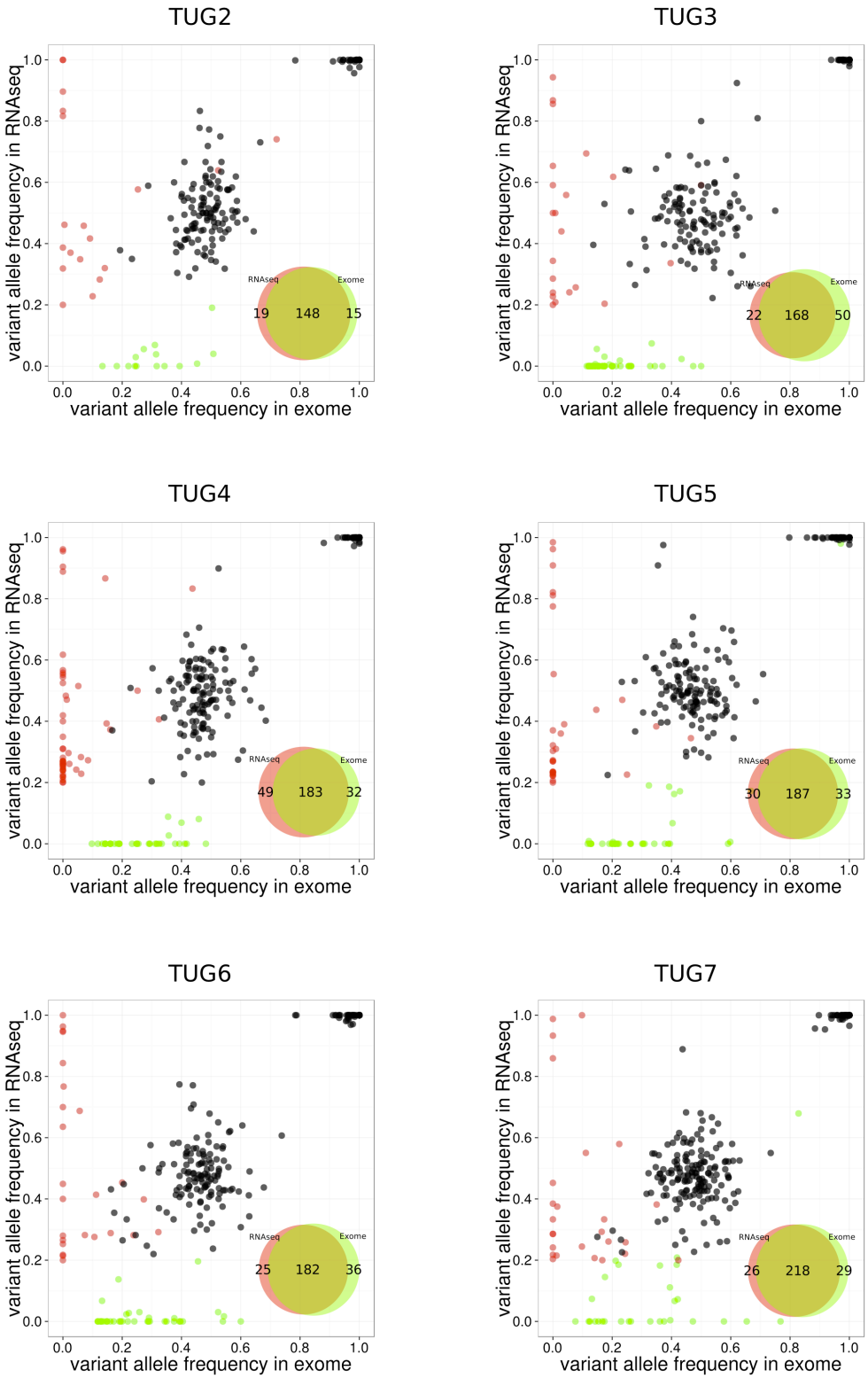

TUG2
variant allele frequency in RNAseq
variant allele frequency in exome
RNAseq
Exome
19
148
15
TUG3
22
168
50
TUG4
49
183
32
TUG5
30
187
33
TUG6
25
182
36
TUG7
26
218
29

Figure S4. Scatter plot of average coverage versus recall ratio per sample.
Recall ratio per sample is calculated as the percentage of Exome-seq SNVs that are called in the RNA-seq as well. Recall ratio 0.3 is assumed as the indicator of a 'good sample' in terms of variant detection.

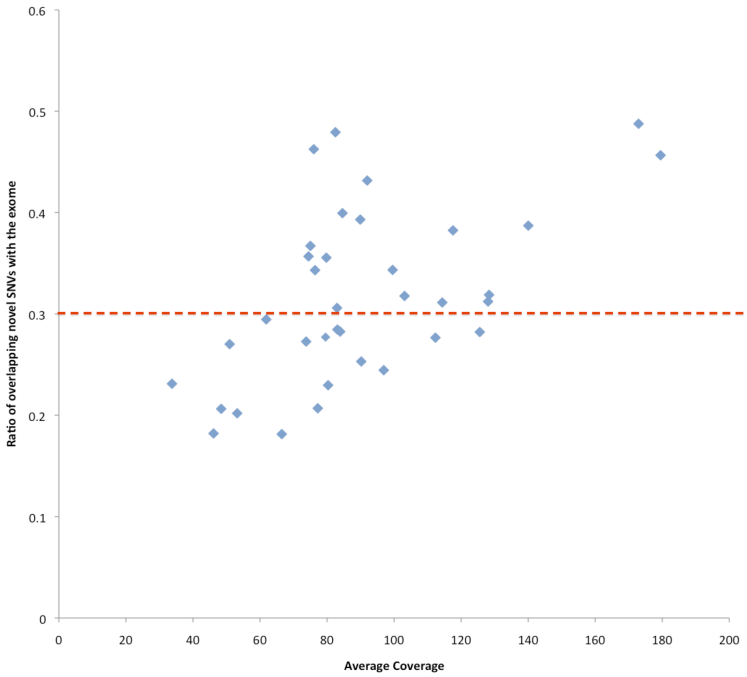
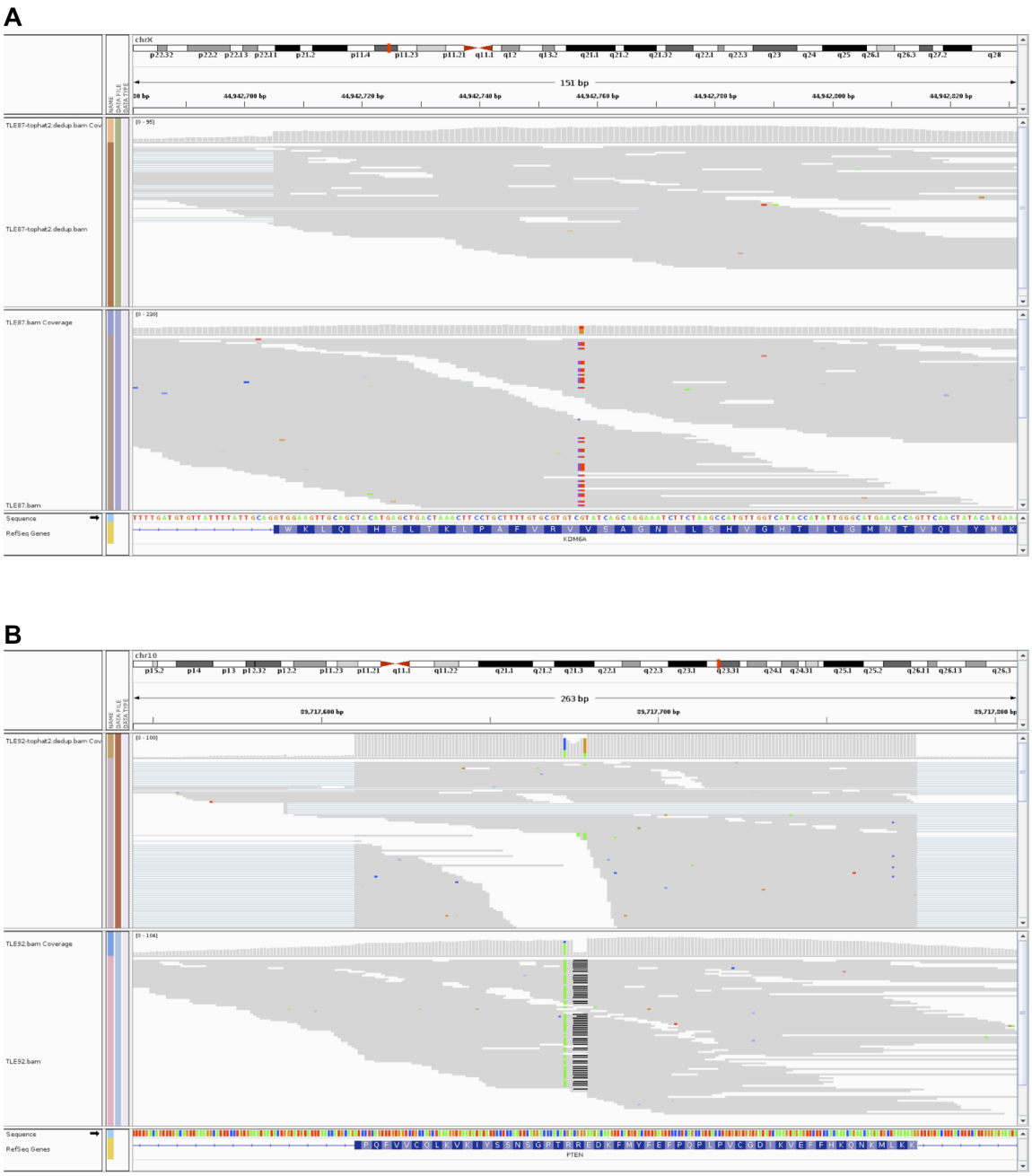
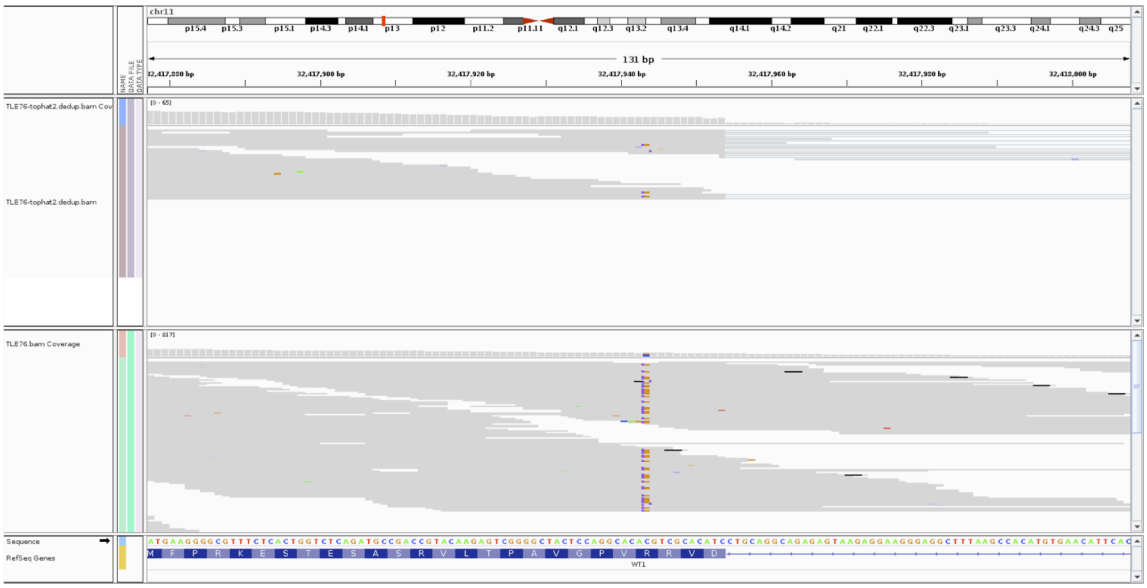


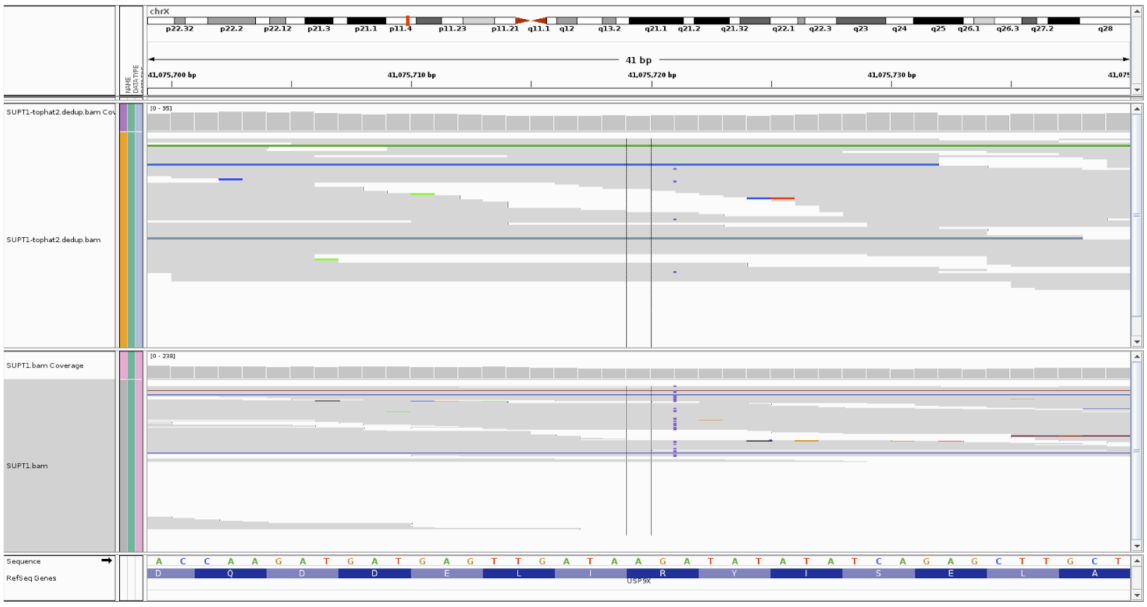
Figure S5. Visualization of the alignments with Exome-seq and RNA-seq for the 5 INDELs that are validated in the DNA of the samples but absent in the RNA-seq alignments.
The Exome-seq and RNA-seq alignment files are visualized using IGV for **A**. *KDM6A* in TLE87, **B**. *PTEN* in TLE92, **C**. *WT1* in TLE76, **D**. *USP9X* in SUPT1, and **E**. *UNC5D* in MOLT4. The exome-seq alignment files (below) have the reads containing the INDEL, whereas RNA-seq alignment files (above) either contain reads with reference only (A, B, and E) or a small portion of reads with INDEL (C and D).



C


chr11
131 bp
TLE76-tophat2.dedup.bam Cov
TLE76-tophat2.dedup.bam
TLE76.bam Coverage
Sequence
RefSeq Genes

D


chrX
41 bp
SUPT1-tophat2.dedup.bam Cov
SUPT1-tophat2.dedup.bam
SUPT1.bam Coverage
SUPT1.bam
Sequence
RefSeq Genes

E

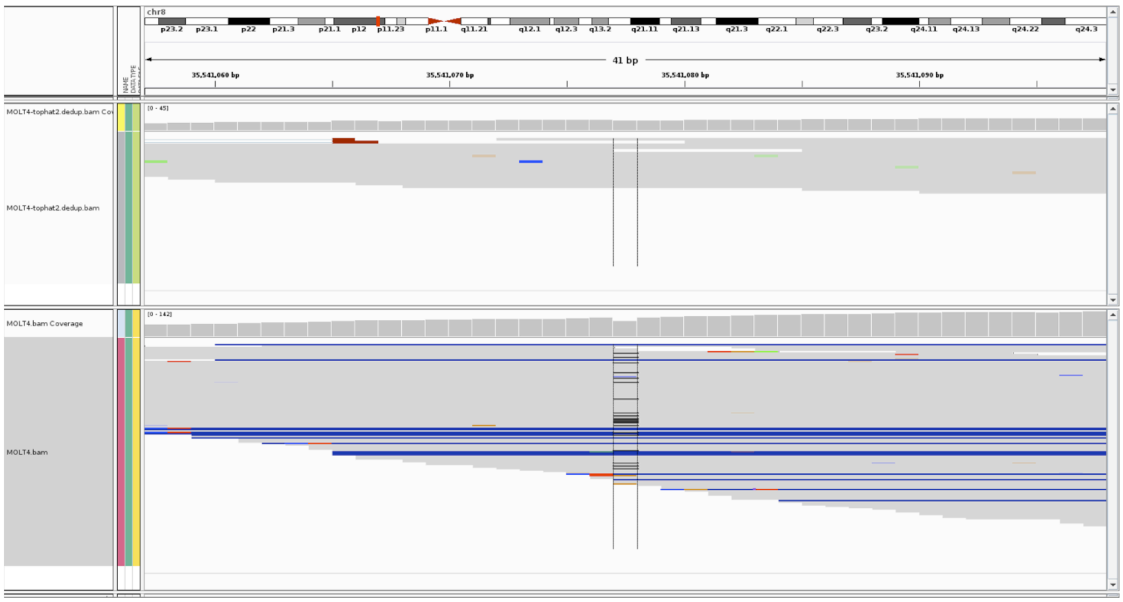
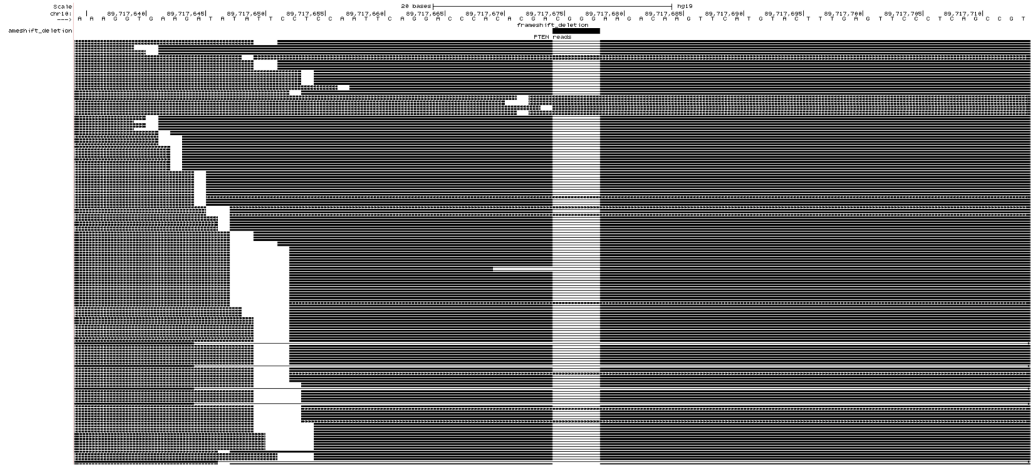
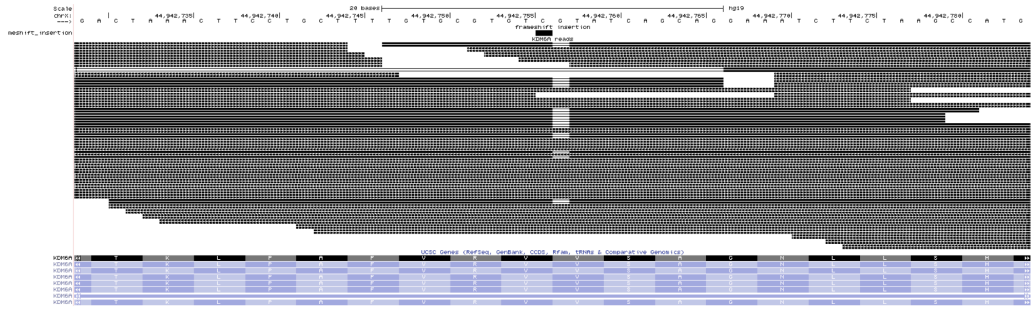

chr8
p23.2 p23.1 p22 p21.3 p21.1 p12 p11.23 p11.1 q11.21 q12.1 q12.3 q13.2 q21.11 q21.13 q21.3 q22.1 q22.3 q23.2 q24.11 q24.13 q24.22 q24.3
41 bp
35,541,060 bp
35,541,070 bp
35,541,080 bp
35,541,090 bp
MOLT4-tophat2.dedup.bam Cov
MOLT4-tophat2.dedup.bam
[0 - 45]
MOLT4.bam Coverage
MOLT4.bam
[0 - 142]

Figure S6. INDELs in TLE92 and TLE87 are detected after mapping with a different aligner.
The screenshots from UCSC genome browser shows **A**. the 4 bp deletion in *PTEN* (note that only a part of the alignment was shown) and **B**. 1bp deletion in *KDM6A*. In both cases BWA transcriptome-only mapping was coupled to BLAT genome mapping. In **C** and **D**, TopHat2 transcriptome-only mapping coupled with BLAT genome mapping was displayed for *PTEN* and KMD6A *INDELs*, respectively.

A



B



C


frameshift_deletion
TLE02_topnat_transcriptome_only
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
Retrocopied Genes V8, Including Pseudogenes

D

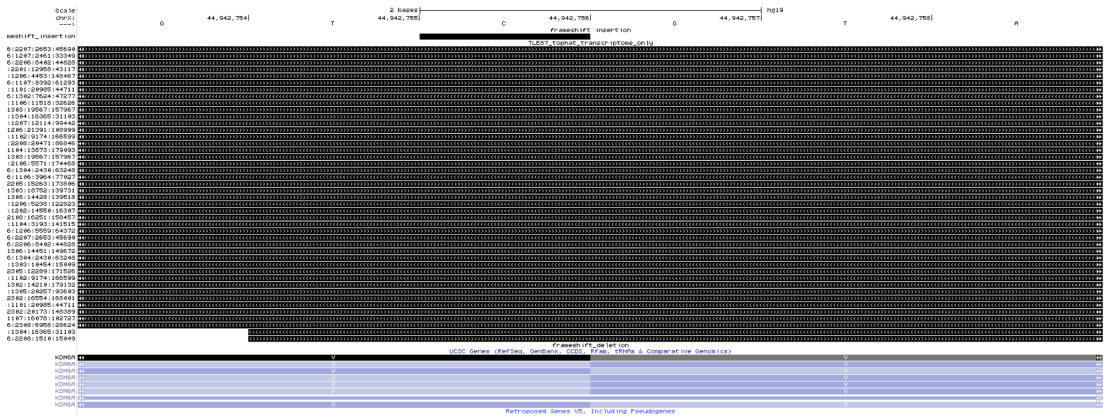

frameshift_insertion
TLE07_topnat_transcriptome_only
frameshift_deletion
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
Retrocopied Genes V8, Including Pseudogenes

Figure S7. Batch effect removal for gene expression profiling.
Multidimensional scaling (MDS) plots before and after batch effect removal. A batch effect was observed whereby samples originating from the same collection center clustered together based on the edgeR normalized gene-by-gene counts (**A**). A similar clustering was observed when the FPKM values per transcript was used (**B**). After fitting a Generalized Linear Model (on the edgeR normalized gene-by-gene counts) accounting for sample collection center, the aberrant clustering of the samples is corrected (**C**).

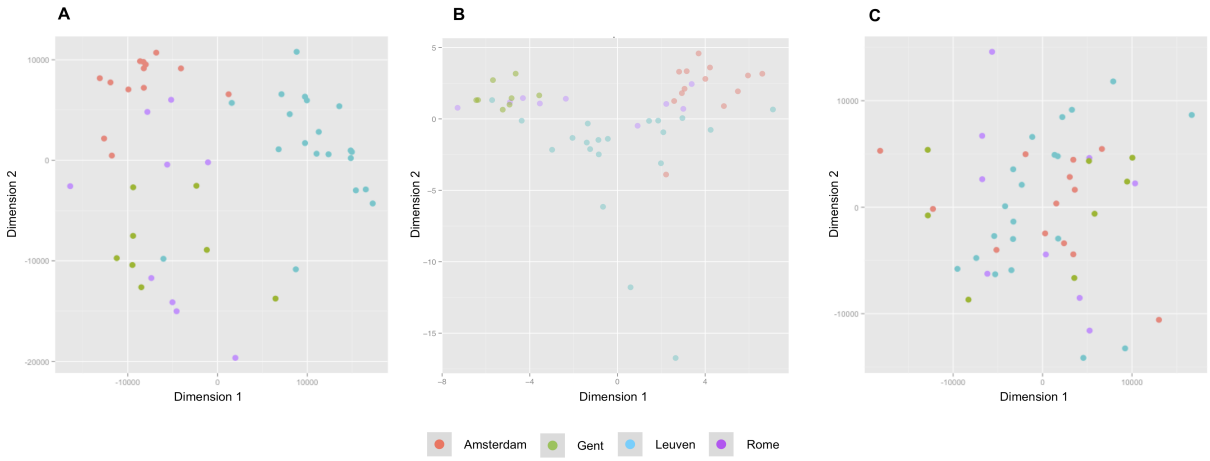
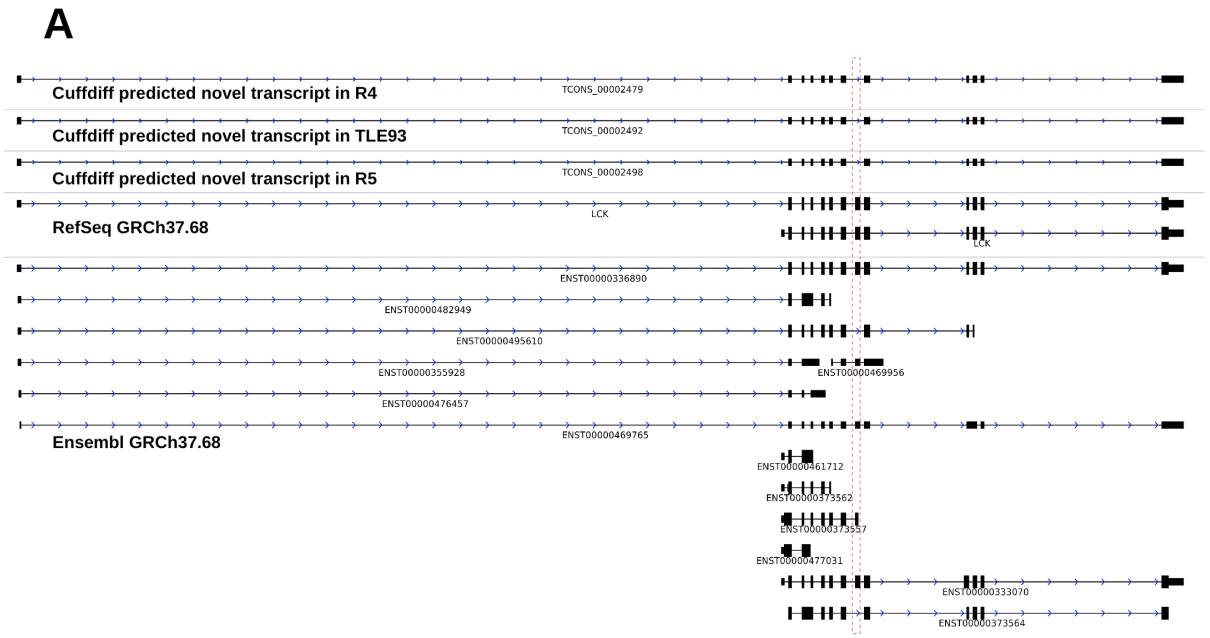
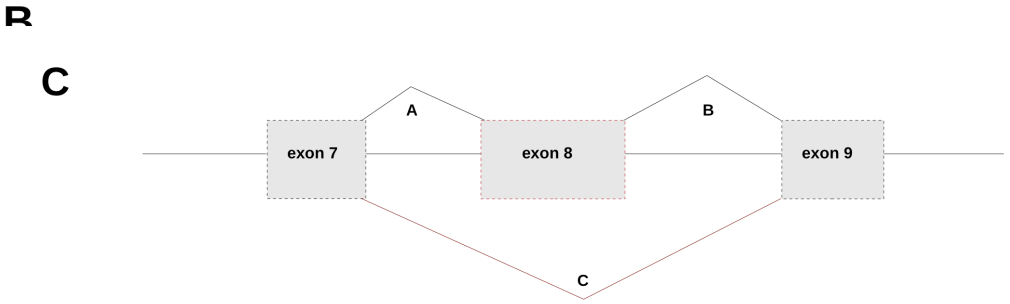


Figure S8. Overview of exon skipping event in *LCK*.
(**A**) Predicted novel transcript of *LCK* aligned with known *LCK* isoforms. Dotted red box indicates the exon-skipping event in the 8th exon (**B**) Sashimi plot detailing the junction supporting the exon skipping event in patient samples R5, R5 and TLE93 with respect to Thymus. (**C**) Schematic representation of the predicted alternative splicing event of *LCK*. The exon skipping ratio (C/A+B+C) of exon 8 of *LCK* in R5, R4, TLE93 are 0.40, 0.47 and 0.20, respectively. (**D**) Schematic overview of *LCK* protein illustrating the spliced out portion without affecting the functional domains.





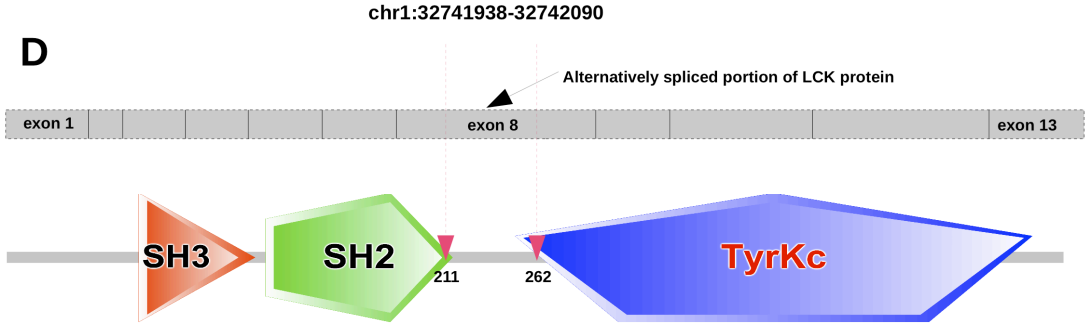


Figure S9. Schematic overview of the SUZ12 exon-skipping event.
(**A**) Schematic representation of the predicted alternative splicing event of SUZ12. The exon skipping ratio (C/A+B+C) of exon 7 of *SUZ12* in R5 is 0.35. (**B**) Schematic overview of SUZ12 protein illustrating the spliced out portion without affecting the functional domains.

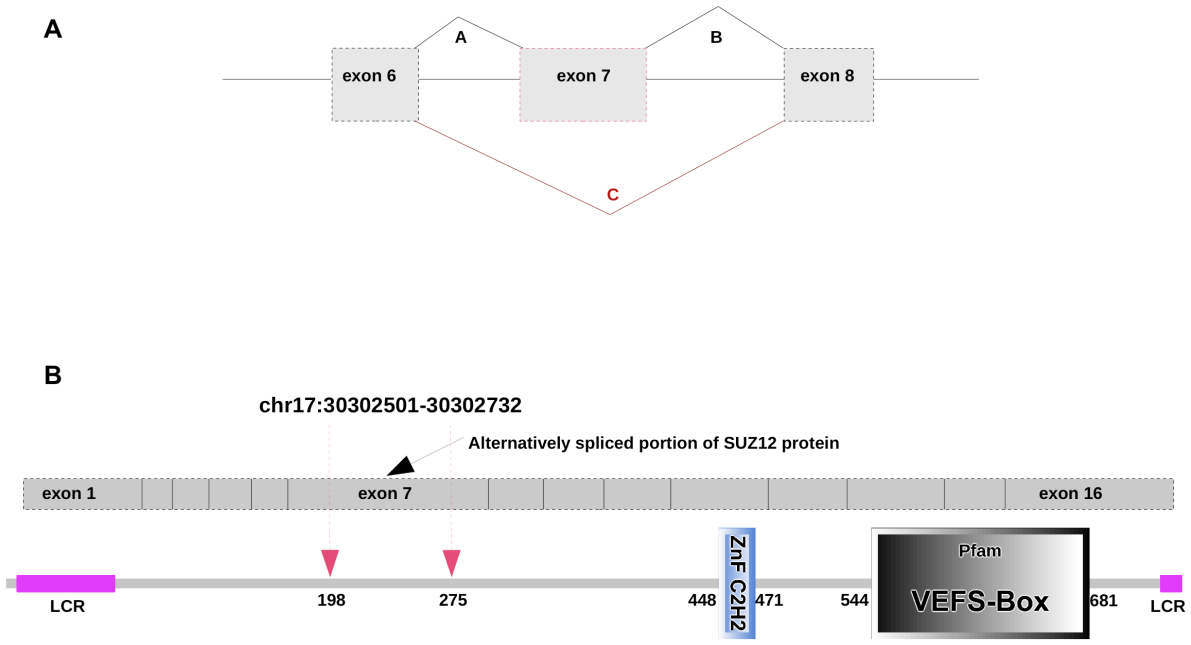


Figure S10. Out-of-frame fusions can have various consequences.
The over or under expression caused by out-of-frame gene fusions are illustrated in the normalized expression heatmap. *CLINT1-MEF2C, HNRP-ZNF219, ZEB1-BMI1* and *AHI1-MYB* fusion are associated with overexpression of *MEF2C, ZNF219, BMI1* and *MYB*; whereas as *TP53-TBC1D3F, PTEN-RNLS, IKZF1-ABCA13* and *CDKN2A-miR31HG* fusions are responsible for the underexpression of *TP53, PTEN, IKZF1* and *CDKN2A*.

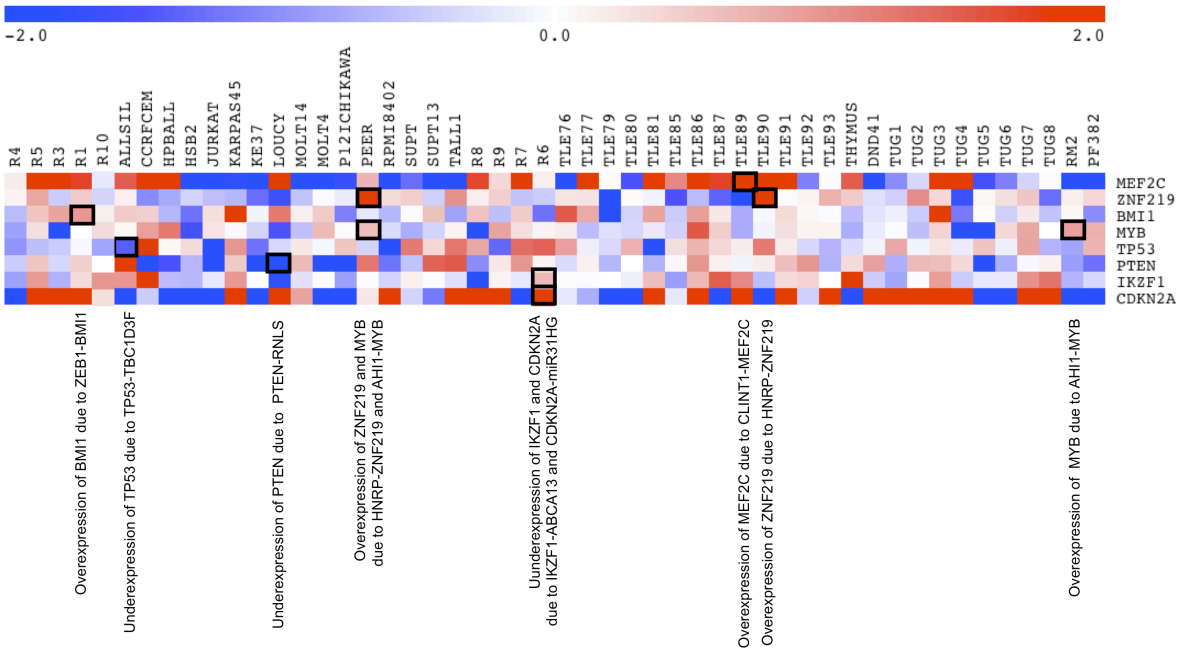


Table S1. (A) Sequencing and mapping statistics, (B) Variant statistics, (C) Fusion statistics
Table S1.A. Sequencing and mapping statistics

	Sample	Total Number of Reads	Mean Target Coverage	2X Coverage (%)	10X Coverage (%)	20X Coverage (%)	30X Coverage (%)
Cell Lines	ALLSIL	58424415	90,2	0,49	0,39	0,35	0,32
	CCRFCEM	62730572	61,86	0,48	0,4	0,34	0,3
	DND41	88001825	179,51	0,52	0,43	0,4	0,38
	HPBALL	88897354	135,8	0,58	0,45	0,4	0,37
	HSB2	40873291	74,47	0,53	0,42	0,37	0,33
	JURKAT	88221222	172,95	0,56	0,46	0,42	0,4
	KARPAS45	65822553	89,9	0,94	0,71	0,53	0,43
	KE37	58459669	79,79	0,92	0,66	0,48	0,39
	LOUCY	69422887	82,5	0,94	0,73	0,56	0,45
	MOLT4	94708631	128,03	0,95	0,78	0,61	0,5
	MOLT14	87567592	145,97	0,57	0,45	0,41	0,38
	P12ICHIKAWA	53998577	76,42	0,91	0,64	0,46	0,37
	PEER	45406860	77,23	0,5	0,36	0,28	0,24
	PF382	83334490	140,04	0,6	0,46	0,41	0,38
	RPMI8402	79017501	125,56	0,59	0,46	0,41	0,38
	SUPT1	49895390	48,42	0,48	0,37	0,3	0,24
	SUPT13	58387261	75,04	0,91	0,61	0,44	0,36
	TALL1	42768080	80,32	0,48	0,39	0,34	0,3
Patients	R1	60261677	73,24	0,57	0,45	0,38	0,34
	R3	4360841	7,35	0,19	0,04	0,02	0,01
	R4	53008343	69,74	0,57	0,42	0,34	0,29
	R5	25383022	46,34	0,49	0,36	0,29	0,23
	R6	28409604	48,88	0,51	0,38	0,29	0,23
	R7	85277047	102,01	0,97	0,83	0,67	0,55
	R8	73952998	101,1	0,96	0,79	0,61	0,49
	R9	93048786	82,08	0,96	0,78	0,59	0,48
	R10	101832247	93,62	0,96	0,79	0,62	0,5
	RM2	64790950	90,46	0,57	0,44	0,37	0,33
	TLE76	31257065	46,14	0,84	0,47	0,3	0,23
	TLE77	52836140	66,49	0,8	0,42	0,3	0,24
	TLE79	29413263	33,74	0,91	0,59	0,32	0,19
	TLE80	93907574	128,33	0,85	0,52	0,41	0,36
	TLE81	62090742	76,04	0,95	0,76	0,57	0,44
	TLE85	36800184	50,94	0,93	0,65	0,4	0,26
	TLE86	58813818	84,6	0,94	0,71	0,51	0,4
	TLE87	64363651	91,98	0,93	0,7	0,52	0,41
	TLE89	79024483	117,57	0,93	0,7	0,51	0,42
	TLE90	54605273	83,09	0,94	0,7	0,49	0,37
	TLE91	56011156	82,98	0,91	0,62	0,44	0,35

	TLE92	69228913	103,11	0,9	0,59	0,42	0,34
	TLE93	50392641	73,76	0,9	0,58	0,39	0,3
	TUG1	63614364	96,91	0,56	0,43	0,37	0,32
	TUG2	36014631	53,19	0,52	0,36	0,27	0,21
	TUG3	90673798	112,33	0,6	0,46	0,4	0,36
	TUG4	62401567	83,91	0,59	0,45	0,38	0,33
	TUG5	83629512	114,37	0,59	0,46	0,41	0,37
	TUG6	62199073	79,55	0,57	0,44	0,38	0,33
	TUG7	85454478	99,56	0,61	0,47	0,41	0,37
	TUG8	61922165	68,23	0,63	0,47	0,39	0,34
	Thymus	127342408	128,77	0,97418	0,85528	0,712207	0,60432
	Average	64365211.68	89,09				
	Min	4360841	7,35				
	Max	127342408	179,51				

Table S1.B. Variant Statistics

		Initial Variant Calling		Custom Variant Filtering	
	Sample	SNVs	INDELs	SNVs	INDELs
Cell lines	ALLSIL	10464	63	555	25
	CCRFCEM	10464	120	1201	56
	DND41	9980	136	766	26
	HPBALL	8505	57	19	8
	HSB2	7033	78	471	23
	JURKAT	11725	169	1836	59
	KARPAS45	14392	84	1532	6
	KE37	9883	103	62	4
	LOUCY	12761	121	65	12
	MOLT4	14612	126	256	18
	MOLT14	8086	50	73	3
	P12ICHIKAWA	9853	76	101	4
	PEER	5161	70	64	3
	PF382	8856	44	511	4
	RPMI8402	7572	42	116	4
	SUPT1	7589	92	979	5
	SUPT13	8916	92	67	42
	TALL1	5873	46	75	3
Patient sample	R1	7072	69	63	3
	R3	291	7	2	0
	R4	5710	81	49	2
	R5	4420	54	46	5
	R6	4485	61	43	4
	R7	14592	107	109	9
	R8	13022	102	75	8
	R9	13550	117	91	4
	R10	14760	123	74	5
	RM2	7505	51	83	4
	TLE76	4797	54	36	1
	TLE77	5753	65	35	0
	TLE79	7743	85	23	0
	TLE80	8905	103	42	8
	TLE81	14038	117	47	2
	TLE85	9909	87	27	5
	TLE86	12546	101	64	7
	TLE87	12715	109	52	7
	TLE89	12108	95	64	7
	TLE90	10190	85	54	1
	TLE91	8921	85	45	5
	TLE92	8469	87	44	4
	TLE93	7922	91	34	6

	TUG1	7143	36	5	3
	TUG2	5013	26	42	2
	TUG3	7983	46	9	2
	TUG4	8217	52	11	3
	TUG5	8493	54	9	0
	TUG6	7765	43	15	3
	TUG7	8355	53	12	4
	TUG8	8847	62	349	11
Total	Sum	442964	3877	10403	430
	Average	9040,1	79,1	212,3	8,8
	Median	8505	81	63	4
	Max	14760	169	1836	59
	Min	291	7	2	0
Patients only	Sum	271239	2308	1654	125
	Average	8749,6	74,5	53,4	4,0
	Median	8355	81	44	4
	Max	14760	123	349	11
	Min	291	7	2	0

Table S1.C. Fusions statistics

	Sample	# of fusions before filtering	# of fusions after defuse filtering	# of fusions after custom filtering
Cell Lines	ALLSIL	585	80	17
	CCRFCEM	512	40	5
	DND41	796	37	3
	HPBALL	1963	113	19
	HSB2	1136	99	10
	JURKAT	1263	68	10
	KARPAS45	833	71	10
	KE37	685	44	6
	LOUCY	820	91	10
	MOLT4	704	19	0
	MOLT14	1381	104	15
	P12ICHIKAWA	690	37	6
	PEER	563	19	4
	PF382	1378	80	11
	RPMI8402	2919	239	42
	SUPT1	485	84	24
	SUPT13	576	19	4
	TALL1	466	21	4
Patients	R1	770	59	7
	R3	209	4	0
	R4	632	43	8
	R5	316	21	2
	R6	431	21	7
	R7	972	44	9
	R8	824	33	4
	R9	1283	71	5
	R10	1686	74	4
	RM2	710	37	4
	TLE76	692	43	9
	TLE77	724	26	4
	TLE79	1101	17	4
	TLE80	1255	53	11
	TLE81	669	38	5
	TLE85	517	17	3
	TLE86	580	15	5
	TLE87	675	14	5
	TLE89	1052	46	12
	TLE90	1269	72	13

	TLE91	792	37	7
	TLE92	1202	69	10
	TLE93	829	35	2
	TUG1	552	33	3
	TUG2	341	12	1
	TUG3	1014	38	8
	TUG4	653	18	3
	TUG5	1019	55	6
	TUG6	565	24	6
	TUG7	852	56	4
	TUG8	627	35	0
	Thymus	1633	60	4
Total	Sum	42568	2425	371
	Average	868.73	49.49	7.57
	Median	724	38	6
	Max	2919	239	42
	Min	209	4	0
Patients only	Sum	24813	1160	171
	Average	800.42	37.42	5.52
	Median	724	37	5
	Max	1686	74	13
	Min	209	4	0

Table S2. Samples analyzed in this study

	Sample	RNAseq	Diagnosis exome	Remission exome
Cell Lines	ALLSIL	yes	yes (*)	NA
	CCRFCEM	yes	yes (*)	NA
	DND41	yes	yes (*)	NA
	HPBALL	yes	no	NA
	HSB2	yes	yes (*)	NA
	JURKAT	yes	yes (*)	NA
	KARPAS45	yes	yes (*)	NA
	KE37	yes	yes (*)	NA
	LOUCY	yes	yes (*)	NA
	MOLT4	yes	yes (*)	NA
	MOLT14	yes	no	NA
	P12ICHIKAWA	yes	yes (*)	NA
	PEER	yes	yes (*)	NA
	PF382	yes	yes (*)	NA
	RPMI8402	yes	yes (*)	NA
	SUPT1	yes	yes (*)	NA
	SUPT13	yes	yes (*)	NA
	TALL1	yes	yes (*)	NA
Patient Samples	R1	yes	no	no
	R3	yes	no	no
	R4	yes	no	no
	R5	yes	no	no
	R6	yes	no	no
	R7	yes	no	no
	R8	yes	no	no
	R9	yes	no	no
	R10	yes	no	no
	RM2	yes	no	no
	TLE76	yes	yes (*)	no
	TLE77	yes	yes	no
	TLE79	yes	yes (*)	no
	TLE80	yes	yes (*)	no
	TLE81	yes	yes	no
	TLE85	yes	yes (*)	no
	TLE86	yes	yes	no
	TLE87	yes	yes (*)	no
	TLE89	yes	yes (*)	no
	TLE90	yes	yes (*)	no
	TLE91	yes	yes (*)	no
	TLE92	yes	yes (*)	no
	TLE93	yes	yes	no
	TUG1	yes	yes	yes

	TUG2	yes	yes	no
	TUG3	yes	yes	yes
	TUG4	yes	yes	yes
	TUG5	yes	yes	yes
	TUG6	yes	yes	yes
	TUG7	yes	yes	yes
	TUG8	yes	no	no
	THYMUS	yes	no	no

* published previously in De Keersmaecker K, Atak ZK, Li N, Vicente C, Patchett S, et al. (2013) Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. Nat Genet 45: 186–190. doi:10.1038/ng.2508

Table S3. Comparison of the number of novel SNV and INDELs between RNAseq and Exome-seq

Sample	RNA-seq		Exome-seq				Overlap		
	SNVs	INDELs	SNVs[a]	SNVs[b]	INDELs[c]	INDELs[d]	SNVs	INDELs[e]	INDELs[f]
ALLSIL	3017	29	652	190	52	19	165	2	2
CCRFCEM	3921	49	2923	997	544	148	861	6	11
DND41	1725	24	2543	1328	230	103	1161	5	6
HSB2	961	18	1906	719	281	112	680	7	8
JURKAT	3222	46	4717	2532	517	246	2300	18	19
KARPAS45	3678	19	5795	3345	110	89	2278	2	8
KE37	407	22	647	303	50	41	230	1	5
LOUCY	515	20	530	313	63	50	254	3	7
MOLT4	1025	22	1358	807	123	108	424	1	5
P12ICHIKAWA	469	17	746	352	54	43	256	0	4
PEER	244	5	667	160	38	10	138	2	2
PF382	1082	9	1891	844	63	22	732	1	3
RPMI8402	376	9	794	270	49	19	224	1	1
SUPT13	376	18	651	275	50	36	239	1	5
SUPT1	3055	40	3717	829	371	127	767	6	9
TALL1	257	8	688	182	38	8	158	2	2
TLE76	197	9	714	157	64	37	130	0	8
TLE77	232	8	678	160	77	44	123	0	4
TLE79	327	13	662	223	108	92	153	0	9
TLE80	377	23	712	268	80	54	227	1	8
TLE81	542	22	640	403	70	67	296	0	11
TLE85	394	18	681	251	113	92	184	1	10
TLE86	512	20	691	389	80	68	276	1	12
TLE87	504	22	658	368	77	70	284	1	9
TLE89	464	21	706	369	110	88	270	3	13
TLE90	431	15	738	307	494	471	210	1	8
TLE91	366	15	693	268	102	78	212	2	10
TLE92	344	18	683	262	90	68	217	1	8
TLE93	324	21	638	247	59	46	174	0	10
TUG1	219	10	589	190	78	17	144	0	7
TUG2	196	9	718	159	131	26	145	1	5
TUG3	277	11	593	214	69	23	164	0	9
TUG4	312	9	637	212	139	39	180	1	8
TUG5	287	8	604	217	109	34	188	0	7
TUG6	293	10	646	214	124	34	179	4	8
TUG7	296	14	629	243	122	35	216	4	11

a: SAMTools - default parameters ; b: SAMTools - default parameters & RNA-seq coverage >20; c: DINDEL - minimum depth:15 & minimum variant allele frequency:0.15; d: DINDEL - minimum depth:15 & minimum variant allele frequency:0.15 & RNA-seq coverage > 3 ; e: Overlap of RNA-seq INDELs with Exome-seq INDELs (d); f: Overlap of RNA-seq INDELs with unfiltered Exome-seq INDELs

Table S4. Validated INDELs from the Exome-seq

Sample	Gene	Position	Reference/Variant	Notes
TLE87	KDM6A	chrX: 44942756	C/CAT	No variant reads, BWA+BLAT predicts a deletion in the adjacent base
TLE92	PTEN	chr10: 89717678	ACGGG/A	No variant reads, BWA+BLAT mapping predicts the indel
TLE76	WT1	chr11: 32417942	A/AG	There are variant reads in RNAseq, however the INDEL falls below the detection limit in variant calling step
TLE80	PHF6	chrX: 133551317	TC/T	Detected in RNA-seq
ALLSIL	NOTCH1	chr9: 139390772	G/GCC	Detected in RNA-seq
SUPT1	USP9X	chrX:41075721	A/AT	There are variant reads in RNAseq, however the INDEL falls below the detection limit in variant calling step
DND41	PHF6	chrX: 133511668	G/GA	Detected in RNA-seq
DND41	HMCN1	chr1:186017944	G/GA	There are no reads in RNAseq for this gene
MOLT4	UNC5D	chr8:35541077	AA/A	No variant reads
CCRFCEM	CT47B1	chrX:120008754	GG/G	There are no reads in RNAseq for this gene

Table S5. Mutations detected in 213 genes.
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Table S6. IPA on 213 candidate genes. Note that this enrichment was obtained using all "expressed genes" from the RNA-seq as background gene set to avoid any bias to expressed genes, because mutations can only be detected in expressed genes when using RNA-seq.
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Table S7. ENDEAVOUR results on 213 genes
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Table S8. ATEs identified in known T-ALL drivers

Gene Name	Locus	Sample1	Sample2	FPKMx	FPKMy	log2(FPKMy/FPKMx)	p-value	q-value	Test source
SUZ12	chr17:30264036-30328064	R3	Thymus	0	439979	inf	5.00E-050	0.00994538	cds_exp.diff
MYB	chr6:135502452-135540311	TUG8	Thymus	0	137848	inf	0.00005	0.0413568	cds_exp.diff
RPL10	chrX:153618314-153650065	TUG2	Thymus	0.245318	0	-inf	0.0001	0.025637	cds_exp.diff
TAL1	chr1:47681922-47697892	TLE92	Thymus	577172	0.390812	-720638	0.0003	0.0454118	gene_exp.diff
FLT3	chr13:28577410-28674729	TLE91	Thymus	282925	202236	-712824	0.00005	0.0109118	gene_exp.diff
NKX2-1,SFTA3	chr14:36942411-36992221	R5	Thymus	0	0.715826	inf	0.00005	0.00573353	gene_exp.diff
NKX2-1,SFTA3	chr14:36942411-36992221	TLE90	Thymus	138368	0.803834	-74274	0.00005	0.0155353	gene_exp.diff
NKX2-1,SFTA3	chr14:36942411-36992221	TLE85	Thymus	100.34	0.847136	-688808	0.00005	0.0151664	gene_exp.diff
NKX2-1,SFTA3	chr14:36942411-36992221	R6	Thymus	0	0.777489	inf	0.00005	0.00444771	gene_exp.diff
NKX2-1,SFTA3	chr14:36942411-36992221	R4	Thymus	0	0.651055	inf	0.00005	0.00691144	gene_exp.diff
NKX2-1,SFTA3	chr14:36942411-36992221	TUG6	Thymus	0	0.754103	inf	0.0001	0.0103673	gene_exp.diff
NKX2-1,SFTA3	chr14:36942411-36992221	R1	Thymus	0	0.814263	inf	0.00005	0.0114146	gene_exp.diff
NKX2-1,SFTA3	chr14:36942411-36992221	TUG5	Thymus	0	0.754974	inf	0.00005	0.00669218	gene_exp.diff
NKX2-1,SFTA3	chr14:36942411-36992221	TUG8	Thymus	0	0.72586	inf	0.00005	0.0132969	gene_exp.diff
NKX2-1,SFTA3	chr14:36942411-36992221	TUG7	Thymus	0	0.835087	inf	0.00015	0.0175297	gene_exp.diff
BCL11B	chr14:99635185-99737861	R3	Thymus	0.286987	205988	616543	0.00185	0.0375529	gene_exp.diff
BCL11B	chr14:99635185-99737861	TLE89	Thymus	0.815174	581135	615562	0.0002	0.0349239	gene_exp.diff
NKX2-1,SFTA3	chr14:36942411-36992221	TUG4	Thymus	0	0.855147	inf	0.00025	0.0367317	gene_exp.diff
NKX2-1,SFTA3	chr14:36942411-36992221	TUG1	Thymus	197343	0.785448	-797297	0.00005	0.0118491	gene_exp.diff
IL7R	chr5:35852796-35879705	R3	Thymus	0.971655	255157	47148	0.00175	0.0360998	gene_exp.diff
LCK	chr1:32716839-32751766	R4	Thymus	191491	0	-inf	0.00005	0.0496942	isoform_exp.diff
LCK	chr1:32716839-32751766	TLE93	Thymus	197045	0	-inf	0.00005	0.0341198	isoform_exp.diff
LCK	chr1:32716839-32751766	R5	Thymus	844094	0	-inf	0.00005	0.0439184	isoform_exp.diff
SUZ12	chr17:30264036-30328064	R3	Thymus	0	439979	inf	0.00005	0.0185729	isoform_exp.diff
SUZ12	chr17:30264036-30328064	R5	Thymus	276409	0	-inf	0.00005	0.0439184	isoform_exp.diff
RUNX1	chr21:36160097-37357047	R5	Thymus	0	135643	inf	0.00005	0.0439184	isoform_exp.diff

RUNX1	chr21:36160097-37357047	TLE93	Thymus	0	127977	inf	0.00005	0.0341198	isoform_exp.diff
SETD2	chr3:47057918-47205457	TUG5	Thymus	685273	0	-inf	0.00005	0.0395056	isoform_exp.diff
SETD2	chr3:47057918-47205457	TUG7	Thymus	495803	0	-inf	0.00005	0.0399164	isoform_exp.diff
RPL10	chrX:153618314-153650065	TLE76	Thymus	0.254194	0.0484117	-23925	0.00005	0.0304595	isoform_exp.diff
RPL10	chrX:153618314-153650065	TLE90	Thymus	0.871304	0	-inf	0.00005	0.0456243	isoform_exp.diff
RPL10	chrX:153618314-153650065	TLE76	Thymus	0.254194	0.0484117	-23925	0.00005	0.0225747	tss_group_exp.diff
FLT3	chr13:28577410-28674729	TLE91	Thymus	282767	197891	-715876	0.00005	0.0249612	tss_group_exp.diff

Table S9. Fusions detected in 49 samples and the Thymus
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Table S10. Annotation of fusions with Pegasus

	Sample	FrameShift	Inframe	Unannotated	Total
Cell lines	ALLSIL	7	2	8	17
	CCRFCEM	3	0	2	5
	DND41	1	0	2	3
	HPBALL	7	2	10	19
	HSB2	4	0	6	10
	JURKAT	5	0	5	10
	KARPAS45	3	1	6	10
	KE37	1	1	4	6
	LOUCY	2	0	8	10
	MOLT14	7	0	8	15
	MOLT4	0	0	0	0
	P12ICHIKAWA	2	0	4	6
	PEER	1	0	3	4
	PF382	7	0	4	11
	RPMI8402	12	4	26	42
	SUPT1	9	4	11	24
	SUPT13	1	0	3	4
	TALL1	1	1	2	4
patient samples	R1	3	0	4	7
	R3	0	0	0	0
	R4	3	1	4	8
	R5	0	2	0	2
	R6	3	0	4	7
	R7	2	2	5	9
	R8	1	0	3	4
	R9	3	0	4	7
	R10	1	0	3	4
	RM2	1	0	3	4
	TLE76	4	0	5	9
	TLE77	2	1	1	4
	TLE79	2	1	1	4
	TLE80	3	0	8	11
	TLE81	1	0	4	5
	TLE85	0	0	3	3
	TLE86	1	0	4	5
	TLE87	3	0	2	5
	TLE89	6	0	6	12
	TLE90	3	1	9	13

	TLE91	3	0	4	7
	TLE92	4	2	4	10
	TLE93	0	0	3	3
	TUG1	0	0	3	3
	TUG2	0	0	1	1
	TUG3	2	1	5	8
	TUG4	0	1	2	3
	TUG5	1	0	5	6
	TUG6	1	0	5	6
	TUG7	0	0	5	5
	TUG8	0	0	0	0
	Thymus	1	1	2	4

Table S11. Patient characteristics
The table is not included in the thesis due to space constrains and can be obtained through the published article.

Table S12. Novel Fusion Transcript validated by RT-PCR and Sanger sequencing
The table is not included in the thesis due to space constrains and can be obtained through the published article.

CHAPTER IV: DISCUSSION

The cancer genomics field gained unprecedented traction with the introduction of next generation sequencing technologies. These sequencing technologies have enabled the inquiry of every base pair in a cancer genome, making it possible to obtain a comprehensive list of genomic changes that occurred during the transformation from a normal to a cancerous cell. Large-scale whole genome and exome sequencing based studies across a variety of cancer types broadened our knowledge on genes that are associated with cancer via genomic aberrations. With every sequenced cancer genome, we are getting one step closer to understand the mechanisms that govern these malignancies.

In this thesis we have demonstrated the use of next generation sequencing technology for driver genomic aberration discovery in T-cell acute lymphoblastic leukemia (T-ALL) by using different sequencing techniques. In-depth analysis of targeted sequencing, exome sequencing and transcriptome sequencing studies have yielded several insights into underlying mechanisms of this cancer. In each project, bioinformatics pipelines were constructed, optimized and validated for obtaining the most accurate set of predictions. Each prediction set was evaluated for recapitulating the current body of knowledge on T-ALL; and the novel biological findings were assessed and further validated experimentally.

In *Paper I*, we have analyzed the targeted sequencing that involved sequencing of 58 cancer genes and 39 candidate genes in 15 T-ALL patient samples and 18 cell lines. We have performed a comparative analysis on the aligners and variant callers, including the companion software gsMapper, in the context of cancer mutation discovery. We have demonstrated that gsMapper, which is often the preferred method of analysis due to ease of use, is among the top performers in terms of sensitivity and specificity of SNV predictions. Moreover, we have showed that better accuracies can be achieved with the combination of other aligners and variant callers, and eventually we could identify SNVs with 95% sensitivity and 93% specificity with the pipeline consisting of two aligners BWA-SW and SSAHA2, and the variant caller Atlas-SNP2.

The sequencing experiment resulted in identification of known mutations in known cancer genes, as 58 out of 97 targeted genes were implicated cancer. On the other hand, we had the opportunity to discover novel driver events in the 38 genes that are included in the screen due to their potential role as tumor suppressors and oncogenes. We selected the candidate event based on their frequency across the patient and cell line cohort, and Sanger sequencing was performed to validate the mutations, and to assess their somatic status and recurrence across larger patient cohorts. Eventually, we observed rare somatic mutations in *SPRY3*, *JAK3* and *TET1*.

The initial motivation behind this study was to assess the feasibility of such a targeted approach for the discovery of novel drivers in large patient cohorts. However, the pace of advancement in the NGS field obviated the use of targeted approaches for discovery due to dropping costs of exome and whole genome sequencing. Instead, targeted sequencing is now widely used for the validation step of large scale sequencing projects. The mutations identified with exome or whole genome sequencing are re-sequenced for validation [46,99,103,218-221]; or in another cohort of samples to assess the mutation frequencies [218,222,223]. Thus we believe this study still holds its significance since the pipelines described here would be applicable to other re-sequencing studies using the Roche/454 platform or other platforms producing long sequence reads. Furthermore, when catalogues of mutations reach saturation, mainly targeted sequencing will be used in a clinical setting for diagnostics.

In *Paper II*, we have explored the mutational landscape of the coding genome via exome sequencing of 67 T-ALL patient samples with 39 matched remissions and 17 cell lines. The availability of the matched remission samples allowed us to find true somatic mutations and to analyze the recurrence on the additional diagnosis-only samples (28) and cell lines. After an iterative step of pipeline optimization and validation, we have identified high quality somatic mutations from the matched diagnosis-remission samples and high quality variations from the remaining diagnosis-only cases and cell lines. We observed marked differences between the pediatric and adult cases both in the number of mutations and in the mutation spectra. Adult T-ALLs harbored 2.7 fold more protein-altering somatic mutations compared to the pediatric cases, consistent with the observations in other tumors [5]. Overall, the average number of somatic protein altering mutations per sample was 12.2, similar to other liquid tumors [5]. However there were vast differences between samples with six having less than 5 somatic protein altering mutations, and three samples having more than 25 somatic protein altering INDELs. 5/6 samples with low mutational load had normal karyotypes resembling other hematopoietic cancers with stable genomes and normal mutation rates [6]. On the other hand, samples with excessive number of INDELs and mutations indicate an elevated mutations rate. Thus the results from our exome-seq data remain inconclusive whether genomic instability is an essential enabling characteristic for T-ALL tumorigenesis. Various

bioinformatics tools and databases were employed to annotate and filter the predicted variants based on the functional consequence on the protein, mutation significance and frequency across the sample cohort. The final candidate driver list contained 15 genes that included key players consistent with the T-ALL literature and novel genes highlighting novel processes that are disrupted in the T-ALL pathogenesis. One of the key findings was the implication of ribosome mutations in tumorigenesis via mutations in *RPL10* and *RPL5,* which encode for parts of the 60S ribosomal complex. The observed mutations in *RPL10*, which occurred on a mutational hotspot on 14/15 of cases, were shown to cause defects in the ribosome in yeast and lymphoid cells using the functional validation studies conducted by the Laboratory of Molecular Biology of Leukemia and our collaborator Arlen Johnson. Another key finding were mutations observed in *CNOT3*, which is involved in transcriptional and post-transcriptional control of gene expression. Recurrent loss of function mutations in *CNOT3* have led to further investigation of this gene in the context of cancer, and functional studies in the *Drosophila* eye cancer model proposed a tumor-suppressors role for this gene.
Neither of these genes had been associated with T-ALL or any other human cancer previously. By following a large-scale sequencing approach in the coding genome, and implementing optimized analysis strategies we have identified mutations not only in known T-ALL drivers but also in genes and processes not involved with any other cancer. Moreover, the analysis pipelines we have assembled are being used in subsequent sequencing projects and an interface (http://lcbmart.aertslab.org/) we have constructed on the BioMart [224] software facilitated easy access to the analysis results and is being updated with the addition of new sequencing data.

In *Paper III*, transcriptomic landscapes of T-ALLs were explored with the analysis of 31 T-ALL patient samples and 18 cell lines that were sequenced with RNA-seq. This study proved to be the most computationally challenging among the three projects, owing to the diverse range of aberrations that could be discovered. We have measured gene expression levels and identified mutations (SNV and INDEL), intragenic fusions and alternative transcript events after addressing key issues in the analysis. First, we have addressed the normalization and batch effect problem to obtain accurate levels of gene expression data, and validated our results against microarray data. We then successfully used the gene expression data for subtype analysis using known marker genes. Next, we addressed a mapping problem in the RNA-seq data for mutation calling, made use of matched exome data in a subset of the cases to assess different mapping strategies, and optimized our mapping and variant calling strategy accordingly. The final mapping strategy resulted in high quality SNV predictions, recovering 32% of the SNVs in matched exomes. Third, we identified fusion events in the RNA-seq data, implemented additional filters and annotations to better characterize these events. And lastly, we detected alternative transcript events. We have used the T-ALL literature to assess the validity of the predictions and obtained driver genes and events. Reassuringly, we found known

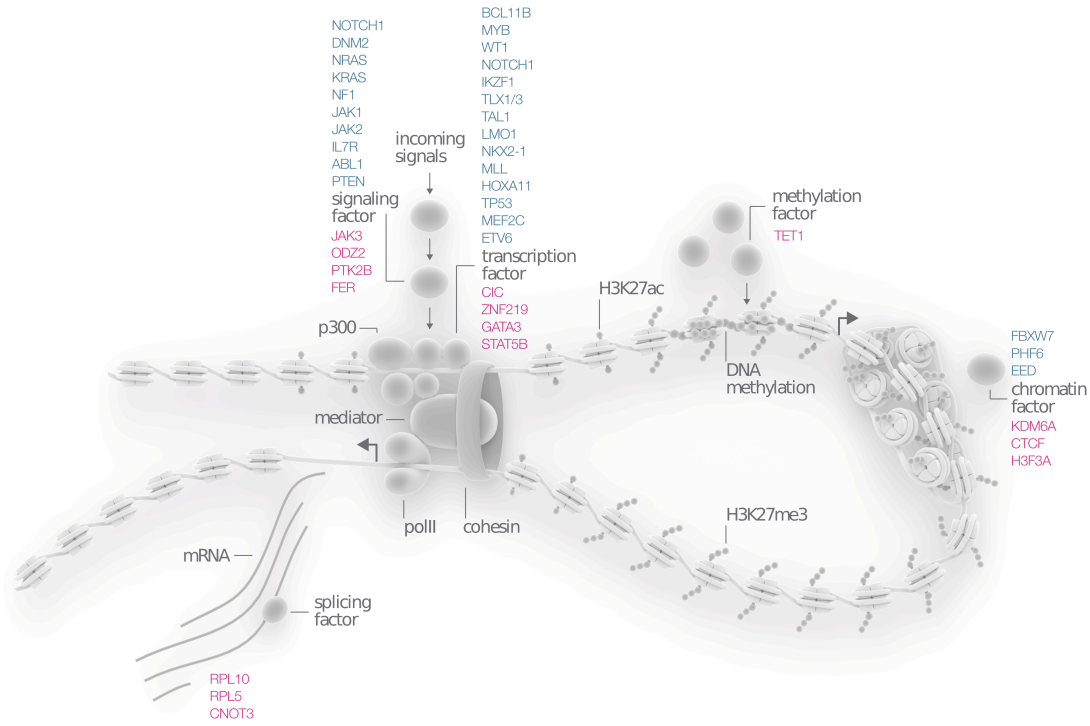
driver events in these samples such as *STIL-TAL1* fusion, which results in *TAL1* overexpression and associated with the *TAL/LMO* subgroup, or *NOTCH1* mutations. For the identification of novel drivers, we have performed a gene prioritization technique on the mutated genes to select genes that are 'functionally similar' to known T-ALL drivers. And in the case of fusion events and ATEs, we have evaluated the events observed in known driver genes and mined for novel driver events in those. Eventually, we identified candidate drivers genes affected by mutations including *CIC*, *H3F3A, PTK2B, STAT5B, ANKRD1, HADHA and DOCK2*; and we identified and validated two novel oncogenic fusions *SSBP2-FER* and *TPM3-JAK2*. ATE discovery resulted in the identification of exon skipping events in *SUZ12* and *LCK*, with the former being validated by PCR.
Our transcriptome sequencing study marks one of the few published studies embarking on complete characterization of variations identifiable from RNA-seq data. We have demonstrated the power of this approach for novel driver detection, however there are still limitations and challenges to address. A first issue concerns the mapping of RNA-seq data for mutation detection. Although we could obtain high quality SNV predictions with our combined mapping strategy, INDEL detection was not as successful. A more refined mapping strategy might be necessary for INDEL detection in RNA-seq data. Second is the lack of systematic ways of assessing gene fusion predictions. Although fusions are identified accurately with existing methods, mathematical models predicting the functional consequences of fusion predictions is lacking and predictions require individual assessment to evaluate their potential effect. A similar problem exists for ATE predictions also, as current methods are limited to prediction algorithms. However, this issue is further complicated by the fact that ATEs have not been causally implicated in cancer although a number of events are associated with specific cancer types [116].

In this thesis, we identified novel driver genes and events in T-ALL with the use of three different NGS approaches.
One general drawback of our sequencing studies, or any sequencing in that matter, is that the driver prediction methods are biased towards specificity because very stringent thresholds are applied to eliminate false positive calls. Lowering these thresholds would indeed increase the sensitivity but it will be in the expense of specificity, thus there is a trade off between these two important aspects of analysis. A sensible way to overcome this conundrum would be to sequence more samples. In these projects, we partially addressed this problem by sequencing different patient cohorts with exome and transcriptome sequencing. And by incorporating the somatic mutation predictions from the exome sequencing project, we could identify additional mutations in genes that were not selected previously due to high stringency filtering. For instance, genes *ANKRD11, CTCF, DOCK2, H3F3A* and *HADHA* had somatic mutations in one patient sample in the exome sequencing study, and had not been identified as driver genes, but with RNA-seq we found

additional cases with mutations in those genes, permitting us to put forward these genes as novel candidates.

We have identified novel drivers fitting into classical cancer genes such as *ODZ2, PTK2B* and *STAT5B*. But we also identified genes with broad functions such as *RPL10, RPL5* and *CNOT3*. Putting the latter mutations in the context of cancer hallmark processes is hard, which was also pointed out by Imielinski *et al* in their analysis of lung cancer genomes [106]. The authors suggested the addition of a new hallmark encompassing epigenetic and RNA regulators. However, changing the perspective from processes to functions can accommodate these mutations. As pointed out in Chapter I, putting these mutations in the context of gene regulation, which actually mediates most of the cancer processes, is revealing since these mutations fit well into the components of this machinery. A similar view has been proposed by Aerts *et al* for the mutational profile of Acute Myeloid Leukemias (AML) [96,225]. Figure 6 depicts the mutations found in our three sequencing projected onto the components of transcriptional machinery as it was done for AML. This points to exciting opportunities for understanding cancer mutations by studying gene regulation.

Figure 6. Frequent mutations in T-ALL can be linked to the regulation of gene expression. Adapted from [225] by somersault18:24 (www.somersault1824.com) Genes in blue represent the known drivers while genes in purple represent the ones that are identified in our three sequencing projects.



OUTLOOK

Cancer driver discovery using NGS technologies will likely remain a central theme until we can identify the full range of genomic aberrations across different cancers. Here I will mention the emerging trends towards achieving this ambitious goal.

The current focus of cancer genomics is the identification of driver oncogenic variants in the coding genome while the variation in the non-coding genome is largely ignored. Almost all the genes that are implicated in cancer are affected by protein-coding mutations, while there is only one gene affected by non-coding mutations in cancer: *TERT* gene with recurrent mutations in the promoter region [117,118]. The reason for this extremely biased focus is purely practical: the effect of a coding mutation can be estimated from the encoded protein, while for a non-coding variant the effect is difficult to estimate if not impossible. However, this trend is deemed to change, as large-scale projects, which are also NGS-based, are effectively improving our understanding of the non-coding genome. The ENCODE project, which aims at cataloging functional elements in the non-coding genome, has already associated 80% of the genome to a biochemical functional, and identified more than 70,000 and 400,000 promoter and enhancer regions in the human genome, respectively [226]. Incorporating this knowledge would finally shift the discovery scale to the whole genome as we will not only identify non-coding mutations but also interpret them.

Another shift of paradigm is on the level of analysis conducted. Current research on cancer genomics is very detailed within the tumor type but very broad across tumor types. In other words, in-depth analysis is performed for each tumor type specifically, but integration of genomic aberrations across different cancers is lacking. A recent initiative is launched for addressing this issue in the Pan-cancer project [227]. This project was initiated by the TCGA consortium and started with the integrative analysis across 12 cancer types. The analysis results from this initiative and the framework laid out by this project is expected to further improve the knowledge base in cancer genomics.

It is undoubtedly clear that NGS will continue to be a driving force for genomic inquiry in the biomedical research and this ultimately will lead to better diagnostic, prognostic and therapeutic strategies in the clinic.

REFERENCES

1. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).
2. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011).
3. Albertini, R. J., Nicklas, J. A., O'Neill, J. P. & Robison, S. H. In vivo somatic mutations in humans: measurement and analysis. *Annu. Rev. Genet.* **24,** 305–326 (1990).
4. Cervantes, R. B., Stringer, J. R., Shao, C., Tischfield, J. A. & Stambrook, P. J. Embryonic stem cells and somatic cells differ in mutation frequency and type. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 3586–3590 (2002).
5. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).
6. Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nat. Rev. Genet.* **13,** 795–806 (2012).
7. Kinzler, K. W. & Vogelstein, B. Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature* **386,** 761–763 (1997).
8. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability--an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.* **11,** 220–228 (2010).
9. Halazonetis, T. D., Gorgoulis, V. G. & Bartek, J. An oncogene-induced DNA damage model for cancer development. *Science* **319,** 1352–1355 (2008).
10. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396,** 643–649 (1998).
11. Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462,** 1005–1010 (2009).
12. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011).
13. Forment, J. V., Kaidi, A. & Jackson, S. P. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nat. Rev. Cancer* **12,** 663–670 (2012).
14. Mitelman, F., Johansson, B. & Mertens, F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat Genet* **36,** 331–334 (2004).
15. Pikor, L., Thu, K., Vucic, E. & Lam, W. The detection and implication of genome instability in cancer - Online First - Springer. *Cancer and Metastasis Reviews* (2013).
16. Vollebergh, M. A., Jonkers, J. & Linn, S. C. Genomic instability in breast and ovarian cancers: translation into clinical predictive biomarkers. *Cell. Mol. Life Sci.* **69,** 223–245 (2012).
17. Thu, K. L. *et al.* Lung adenocarcinoma of never smokers and smokers harbor differential regions of genetic alteration and exhibit different levels of genomic

instability. *PLoS ONE* **7,** e33003 (2012).

18. Jhappan, C., Noonan, F. P. & Merlino, G. Ultraviolet radiation and cutaneous malignant melanoma. *Oncogene* **22,** 3099–3112 (2003).
19. Pfeifer, G. P. *et al.* Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* **21,** 7435–7451 (2002).
20. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
21. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149,** 979–993 (2012).
22. Taylor, B. J. *et al.* DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *eLife* **2,** (2013).
23. Grivennikov, S. I., Greten, F. R. & Karin, M. Immunity, inflammation, and cancer. *Cell* **140,** 883–899 (2010).
24. Colotta, F., Allavena, P., Sica, A., Garlanda, C. & Mantovani, A. Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability. *Carcinogenesis* **30,** 1073–1081 (2009).
25. Gao, S. P. *et al.* Mutations in the EGFR kinase domain mediate STAT3 activation via IL-6 production in human lung adenocarcinomas. *J. Clin. Invest.* **117,** 3846–3856 (2007).
26. Hussain, S. P., Hofseth, L. J. & Harris, C. C. Radical causes of cancer. *Nat. Rev. Cancer* **3,** 276–285 (2003).
27. Bielas, J. H., Loeb, K. R., Rubin, B. P., True, L. D. & Loeb, L. A. Human cancers express a mutator phenotype. *pnas.org*
28. Prior, I. A., Lewis, P. D. & Mattos, C. A Comprehensive Survey of Ras Mutations in Cancer. *Cancer Res.* **72,** 2457–2467 (2012).
29. Davies, M. A. & Samuels, Y. Analysis of the genome to personalize therapy for melanoma. *Oncogene* **29,** 5545–5555 (2010).
30. Jiang, B. H. & Liu, L. Z. PI3K/PTEN signaling in angiogenesis and tumorigenesis. *Adv. Cancer Res.* **102,** 19–65 (2009).
31. Miller, T. W., Rexer, B. N., Garrett, J. T. & Arteaga, C. L. Mutations in the phosphatidylinositol 3-kinase pathway: role in tumor progression and therapeutic implications in breast cancer. *Breast Cancer Res.* **13,** 224 (2011).
32. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
33. Massagué, J., Blain, S. W. & Lo, R. S. TGFbeta signaling in growth control, cancer, and heritable disorders. *Cell* **103,** 295–309 (2000).
34. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321,** 1801–1806 (2008).
35. Petitjean, A. *et al.* Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments

in the IARC TP53 database. *Hum. Mutat.* **28,** 622–629 (2007).

36. Tommasino, M. *et al.* The role of TP53 in Cervical carcinogenesis. *Hum. Mutat.* **21,** 307–312 (2003).
37. Barretina, J. *et al.* Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nat Genet* **42,** 715–721 (2010).
38. Lane, D. P. Cancer. p53, guardian of the genome. *Nature* **358,** 15–16 (1992).
39. HAYFLICK, L. & MOORHEAD, P. S. The serial cultivation of human diploid cell strains. *Exp. Cell Res.* **25,** 585–621 (1961).
40. Shay, J. W. & Bacchetti, S. A survey of telomerase activity in human cancer. *European Journal of Cancer* **33,** 787–791 (1997).
41. Cesare, A. J. & Reddel, R. R. Telomere uncapping and alternative lengthening of telomeres. *Mech. Ageing Dev.* **129,** 99–108 (2008).
42. Bryan, T. M., Marusic, L., Bacchetti, S., Namba, M. & Reddel, R. R. The telomere lengthening mechanism in telomerase-negative immortal human cells does not involve the telomerase RNA subunit. *Human Molecular Genetics* **6,** 921–926 (1997).
43. Parker, M. *et al.* Assessing telomeric DNA content in pediatric cancers using whole-genome sequencing data. *Genome Biol* **13,** R113 (2012).
44. Rak, J. *et al.* Mutant ras oncogenes upregulate VEGF/VPF expression: implications for induction and inhibition of tumor angiogenesis. *Cancer Res.* **55,** 4575–4580 (1995).
45. J R Gnarra, S. Z. M. J. M. J. R. W. A. K. E. P. E. H. O. R. D. K. W. M. L. Post-transcriptional regulation of vascular endothelial growth factor mRNA by the product of the VHL tumor suppressor gene. *Proc. Natl. Acad. Sci. U.S.A.* **93,** 10589 (1996).
46. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).
47. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467,** 1114–1117 (2010).
48. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467,** 1109–1113 (2010).
49. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).
50. Vignot, S. *et al.* Next-Generation Sequencing Reveals High Concordance of Recurrent Somatic Alterations Between Primary Tumor and Metastases From Patients With Non-Small-Cell Lung Cancer. *Journal of Clinical Oncology* **31,** 2167–2172 (2013).
51. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366,** 883–892 (2012).
52. Vakiani, E. *et al.* Comparative genomic analysis of primary versus metastatic colorectal carcinomas. *Journal of Clinical Oncology* **30,** 2956–2962 (2012).
53. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467,** 1109–1113 (2010).

54. Levine, A. J. & Puzio-Kuter, A. M. The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. *Science* **330,** 1340–1344 (2010).
55. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321,** 1807–1812 (2008).
56. Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361,** 1058–1066 (2009).
57. Amary, M. F. *et al.* IDH1 and IDH2 mutations are frequent events in central chondrosarcoma and central and periosteal chondromas but not in other mesenchymal tumours. *J. Pathol.* **224,** 334–343 (2011).
58. Zhang, C., Moore, L. M. & Li, X. IDH1/2 mutations target a key hallmark of cancer by deregulating cellular metabolism in glioma. *Neuro- …* (2013).
59. Vesely, M. D., Kershaw, M. H., Schreiber, R. D. & Smyth, M. J. Natural innate and adaptive immunity to cancer. *Annual review of immunology* **29,** 235–271 (2011).
60. Dunn, G. P., Koebel, C. M. & Schreiber, R. D. Interferons, immunity and cancer immunoediting. *Nat. Rev. Immunol.* **6,** 836–848 (2006).
61. Network, T. C. G. A. R. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).
62. Sanda, T. *et al.* Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell* **22,** 209–221 (2012).
63. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153,** 17–37 (2013).
64. Jones, S. *et al.* Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* **330,** 228–231 (2010).
65. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499,** 43–49 (2013).
66. Love, C. *et al.* The genetic landscape of mutations in Burkitt lymphoma. *Nat Genet* **44,** 1321–1325 (2012).
67. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152,** 1237–1251 (2013).
68. Nagl, N. G., Zweitzig, D. R., Thimmapaya, B., Beck, G. R. & Moran, E. The c-myc gene is a direct target of mammalian SWI/SNF-related complexes during differentiation-associated cell cycle arrest. *Cancer Res.* **66,** 1289–1293 (2006).
69. Meijerink, J. P. P. Genetic rearrangements in relation to immunophenotype and outcome in T-cell acute lymphoblastic leukaemia. *Best Pract Res Clin Haematol* **23,** 307–318 (2010).
70. Aplan, P. D. *et al.* Disruption of the human SCL locus by 'illegitimate' V-(D)-J recombinase activity. *Science* **250,** 1426–1429 (1990).
71. van Vlierberghe, P. *et al.* The cryptic chromosomal deletion del(11)(p12p13) as a new activation mechanism of LMO2 in pediatric T-cell acute

lymphoblastic leukemia. *Blood* **108,** 3520–3529 (2006).
72. Lahortiga, I. *et al.* Duplication of the MYB oncogene in T cell acute lymphoblastic leukemia. *Nat Genet* **39,** 593–595 (2007).
73. Ferrando, A. A. *et al.* Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* **1,** 75–87 (2002).
74. Weng, A. P. *et al.* Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* **306,** 269–271 (2004).
75. Ferrando, A. A. The role of NOTCH1 signaling in T-ALL. *Hematology Am Soc Hematol Educ Program* 353–361 (2009). doi:10.1182/asheducation-2009.1.353
76. van Vlierberghe, P. & Ferrando, A. The molecular basis of T cell acute lymphoblastic leukemia. *J. Clin. Invest.* **122,** 3398–3406 (2012).
77. Hebert, J., Cayuela, J. M., Berkeley, J. & Sigaux, F. Candidate tumor-suppressor genes MTS1 (p16INK4A) and MTS2 (p15INK4B) display frequent homozygous deletions in primary cells from T- but not from B-cell lineage acute lymphoblastic leukemias. *Blood* **84,** 4038–4044 (1994).
78. Dang, C. V. MYC on the Path to Cancer. *Cell* **149,** 22–35 (2012).
79. Erikson, J. *et al.* Deregulation of c-myc by translocation of the alpha-locus of the T-cell receptor in T-cell leukemias. *Science* **232,** 884–886 (1986).
80. Palomero, T., Dominguez, M. & Ferrando, A. A. The role of the PTEN/AKT Pathway in NOTCH1-induced leukemia. *Cell Cycle* **7,** 965–970 (2008).
81. Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481,** 157–163 (2012).
82. Ntziachristos, P. *et al.* Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. *Nat Med* **18,** 298–301 (2012).
83. van Vlierberghe, P. *et al.* PHF6 mutations in T-cell acute lymphoblastic leukemia. *Nat Genet* **42,** 338–342 (2010).
84. Boveri, T. *Zur Frage der Entstehung maligner Tumoren.* (G. Fischer, 1914).
85. Hansemann, D. Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung. *Archiv f. pathol. Anat.* **119,** 299–326 (1890).
86. ROWLEY, J. D. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature* **243,** 290–293 (1973).
87. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300,** 149–152 (1982).
88. Tabin, C. J. *et al.* Mechanism of activation of a human oncogene. *Nature* **300,** 143–149 (1982).
89. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
90. Sjoblom, T. *et al.* The Consensus Coding Sequences of Human Breast and

Colorectal Cancers. *Science* **314,** 268–274 (2006).

91. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113 (2007).
92. Dickson, D. Wellcome funds cancer database. *Nature* (1999).
93. Collins, F. S. & Barker, A. D. Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci. Am.* **296,** 50–57 (2007).
94. International Cancer Genome Consortium *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
95. Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10,** Unit 10.11 (2008).
96. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368,** 2059–2074 (2013).
97. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
98. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486,** 400–404 (2012).
99. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486,** 395–399 (2012).
100. Quesada, V. *et al.* Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet* **44,** 47–52 (2012).
101. Seshagiri, S. *et al.* Recurrent R-spondin fusions in colon cancer. *Nature* **488,** 660–664 (2012).
102. Zhang, J. *et al.* Genetic heterogeneity of diffuse large B-cell lymphoma. *Proceedings of the National Academy of Sciences* **110,** 1398–1403 (2013).
103. Lohr, J. G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proceedings of the National Academy of Sciences* **109,** 3879–3884 (2012).
104. Cancer Genome Atlas Research Network *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).
105. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333,** 1157–1160 (2011).
106. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150,** 1107–1120 (2012).
107. Krauthammer, M. *et al.* Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat Genet* **44,** 1006–1014 (2012).
108. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150,** 251–263 (2012).
109. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
110. Biankin, A. V. *et al.* Pancreatic cancer genomes reveal aberrations in axon

guidance pathway genes. *Nature* **491,** 399–405 (2012).
111. Jones, D. T. W. *et al.* Dissecting the genomic complexity underlying medulloblastoma. *Nature* **488,** 100–105 (2012).
112. Molenaar, J. J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483,** 589–593 (2012).
113. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* **44,** 685–689 (2012).
114. de Keersmaecker, K. *et al.* Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nat Genet* **45,** 186–190 (2013).
115. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7,** 85–97 (2006).
116. Venables, J. P. Aberrant and Alternative Splicing in Cancer. *Cancer Res.* (2004).
117. Horn, S. *et al.* TERT Promoter Mutations in Familial and Sporadic Melanoma. *Science* **339,** 959–961 (2013).
118. Huang, F. W. *et al.* Highly Recurrent TERT Promoter Mutations in Human Melanoma. *Science* **339,** 957–959 (2013).
119. Robison, K. Application of second-generation sequencing to cancer genomics. *Brief. Bioinformatics* **11,** 524–534 (2010).
120. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11,** 685–696 (2010).
121. Rivera, C. M. & Ren, B. Mapping Human Epigenomes. *Cell* **155,** 39–55 (2013).
122. Flicek, P. & Birney, E. Sense from sequence reads: methods for alignment and assembly. *Nature Methods* **6,** S6–S12 (2009).
123. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491–498 (2011).
124. Cabanski, C. R. *et al.* ReQON: a Bioconductor package for recalibrating quality scores from next-generation sequencing data. *BMC Bioinformatics* **13,** 221 (2012).
125. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res* **19,** 1124–1132 (2009).
126. Li, H. Improving SNP discovery by base alignment quality. *Bioinformatics* **27,** 1157–1158 (2011).
127. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25,** 2283–2285 (2009).
128. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12,** 443–451 (2011).
129. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework

for analyzing next-generation DNA sequencing data. *Genome Res* **20,** 1297–1303 (2010).

130. Albers, C. A. *et al.* Dindel: accurate indel calls from short-read data. *Genome Res* **21,** 961–973 (2011).
131. Neuman, J. A., Isakov, O. & Shomron, N. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Brief. Bioinformatics* **14,** 46–55 (2013).
132. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22,** 568–576 (2012).
133. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).
134. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
135. Oshlack, A. & Wakefield, M. J. Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* **4,** 14 (2009).
136. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
137. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11,** 94 (2010).
138. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11,** R106 (2010).
139. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
140. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12,** 480 (2011).
141. Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinformatics* (2012). doi:10.1093/bib/bbs046
142. Li, J. & Tibshirani, R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res* (2011). doi:10.1177/0962280211428386
143. Auer, P. L. & Doerge, R. W. A Two-Stage Poisson Model for Testing RNA-Seq Data. *Stat Appl Genet Mol Biol* **10,** 1–26
144. Zhou, Y.-H., Xia, K. & Wright, F. A. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics* **27,** 2672–2678 (2011).
145. Hardcastle, T. J. & Kelly, K. A. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11,** 422 (2010).

146. Leng, N. *et al.* EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29,** 1035–1043 (2013).
147. Di Yanming, W, S. D., S, C. J. & H, C. J. The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. *Stat Appl Genet Mol Biol* **10,** 1–28 (2011).
148. Zhou, Y.-H., Xia, K. & Wright, F. A. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics* **27,** 2672–2678 (2011).
149. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **12,** 671–682 (2011).
150. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods* **8,** 469–477 (2011).
151. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology* **28,** 503–510 (2010).
152. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27,** 2325–2329 (2011).
153. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28,** 511–515 (2010).
154. Griffith, M. *et al.* Alternative expression analysis by RNA sequencing. *Nature Methods* **7,** 843–847 (2010).
155. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* **7,** 1009–1015 (2010).
156. McPherson, A. *et al.* deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol* **7,** e1001138 (2011).
157. Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38,** e178 (2010).
158. Ge, H. *et al.* FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics* **27,** 1922–1928 (2011).
159. Francis, R. W. *et al.* FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data. *PLoS ONE* **7,** e39987 (2012).
160. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* **12,** R72 (2011).
161. Iyer, M. K., Chinnaiyan, A. M. & Maher, C. A. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* **27,** 2903–2904 (2011).
162. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31,** 213–219 (2013).
163. Roberts, N. D. *et al.* A comparative analysis of algorithms for somatic SNV

detection in cancer. *Bioinformatics* **29,** 2223–2230 (2013).
164. International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
165. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
166. Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22,** 1589–1598 (2012).
167. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
168. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).
169. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).
170. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38,** e164 (2010).
171. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31,** 3812–3814 (2003).
172. Adzhubei, I. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
173. Wong, W. C. *et al.* CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* **27,** 2147–2148 (2011).
174. Kaminker, J. S. *et al.* Distinguishing Cancer-Associated Missense Mutations from Common Polymorphisms.
175. Falgueras, J. *et al.* SeqTrim: a high-throughput pipeline for preprocessing any type of sequence reads. *BMC Bioinformatics* **11,** 38 (2010).
176. Patel, R. K. & Jain, M. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE* **7,** e30619 (2012).
177. Zhou, Q., Su, X., Wang, A., Xu, J. & Ning, K. QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. *PLoS ONE* **8,** e60234 (2013).
178. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18,** 1851–1858 (2008).
179. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
180. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
181. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
182. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large

DNA databases. *Genome Res* **11,** 1725–1729 (2001).

183. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10,** R25 (2009).
184. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).
185. Rumble, S. M. *et al.* SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* **5,** e1000386 (2009).
186. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* **21,** 936–939 (2011).
187. Frith, M. C., Wan, R. & Horton, P. Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic Acids Res* **38,** e100 (2010).
188. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).
189. Liu, C.-M. *et al.* SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics* **28,** 878–879 (2012).
190. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41,** 1061–1067 (2009).
191. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature Methods* **7,** 576–577 (2010).
192. Blom, J. *et al.* Exact and complete short-read alignment to microbial genomes using Graphics Processing Unit programming. *Bioinformatics* **27,** 1351–1358 (2011).
193. Homer, N., Merriman, B. & Nelson, S. F. BFAST: an alignment tool for large scale genome resequencing. *PLoS ONE* **4,** e7767 (2009).
194. Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Res* **12,** 656–664 (2002).
195. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26,** 873–881 (2010).
196. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
197. Quinlan, A. R., Stewart, D. A., Stromberg, M. P. & Marth, G. T. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature Methods* **5,** 179–181 (2008).
198. Misra, S., Agrawal, A., Liao, W.-K. & Choudhary, A. Anatomy of a hash-based long read sequence mapping algorithm for next generation DNA sequencing. *Bioinformatics* **27,** 189–195 (2011).
199. Goya, R. *et al.* SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26,** 730–736 (2010).
200. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
201. Shen, Y. *et al.* A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res* **20,** 273–280 (2010).

202. Valdés-Mas, R., Beà, S., Puente, D. A., López-Otín, C. & Puente, X. S. Estimation of copy number alterations from exome sequencing data. *PLoS ONE* **7,** e51422 (2012).
203. Sathirapongsasuti, J. F. *et al.* Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **27,** 2648–2654 (2011).
204. Amarasinghe, K. C., Li, J. & Halgamuge, S. K. CoNVEX: copy number variation estimation in exome sequencing data using HMM. *BMC Bioinformatics* **14 Suppl 2,** S2 (2013).
205. Fromer, M. *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* **91,** 597–607 (2012).
206. Krumm, N. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res* **22,** 1525–1532 (2012).
207. Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* **8,** 1765–1786 (2013).
208. Delhomme, N., Padioleau, I., Furlong, E. E. & Steinmetz, L. M. easyRNASeq: a bioconductor package for processing RNA-Seq data. *Bioinformatics* **28,** 2532–2533 (2012).
209. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9,** e1003118 (2013).
210. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res* **21,** 2213–2223 (2011).
211. Lund, S. P., Nettleton, D., McCarthy, D. J. & Smyth, G. K. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat Appl Genet Mol Biol* **11,** (2012).
212. Van De Wiel, M. A. *et al.* Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* **14,** 113–128 (2012).
213. Li, Y., Chien, J., Smith, D. I. & Ma, J. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics* **27,** 1708–1710 (2011).
214. Larson, D. E. *et al.* SomaticSniper: Identification of Somatic Point Mutations in Whole Genome Sequencing Data. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btr665
215. Saunders, C. T. *et al.* Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* (2012). doi:10.1093/bioinformatics/bts271
216. Roth, A. *et al.* JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics* **28,** 907–913 (2012).
217. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39,**

e118–e118 (2011).

218. Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486,** 353–360 (2012).
219. Ho, A. S. *et al.* The mutational landscape of adenoid cystic carcinoma. *Nat Genet* **45,** 791–798 (2013).
220. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486,** 405–409 (2012).
221. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149,** 979–993 (2012).
222. Rajala, H. L. M. *et al.* Discovery of somatic STAT5b mutations in large granular lymphocytic leukemia. *Blood* (2013). doi:10.1182/blood-2012-12-474577
223. Wu, G. *et al.* Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. *Nat Genet* **44,** 251–253 (2012).
224. Kasprzyk, A. BioMart: driving a paradigm change in biological data management. *Database (Oxford)* **2011,** bar049 (2011).
225. Aerts, S. & Cools, J. Cancer: Mutations close in on gene regulation. *Nature* **499,** 35–36 (2013).
226. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
227. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45,** 1113–1120 (2013).

RESUME

Personal Information

Place and date of birth	Antalya, Turkey. May 1, 1984
Work address	Center for Human Genetics, Faculty of Medicine, K.U.Leuven Laboratory of Computational Biology Herestraat 49, PO Box 602, 3000 LEUVEN, Belgium

Education

2002- 2007	**B.Sc. Statistics** Middle East Technical University Ankara, Turkey
2007- 2009	**M.Sc. Bioinformatics and Systems Biology** Chalmers Institute of Technology Gothenburg, Sweden Thesis Title: "Discovery of commonly expressed genes in Ewing's Sarcoma and Myxoid Liposarcoma: A meta-analysis of publicly available gene expression data" Advisors: Prof. Olle Nerman and Prof. Pierre Åman
2009 – present	**PhD Biomedical Sciences** KULeuven Leuven, Belgium Project: "Identification of oncogenic mutations in large patient groups by means of nextgeneration sequencing" Supervisors: Prof. Stein Aerts and Prof. Jan Cools

Oral Presentations At Meetings

2013 **ESH-ICMLF 15th International Conference CHRONIC MYELOID LEUKEMIA - Biology and Therapy**
Estoril, Portugal. Invited talk.
"Deciphering T-ALL genomes: sequencing and beyond"

2013 **BeNeLuX Bioinformatics Conference 2013**
Brussels, Belgium. Selected talk.
"Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia"

Poster Presentation At Meetings

2010 **ECCB10 9th European Conference on Computational Biology**
Ghent, Belgium
"Comparison of mapping and variant calling accuracies for next-generation sequence data in relation to coverage depth: a case study in leukemia cell lines"

2011 **EMBL Cancer Genomics**
Heidelberg, Germany
"Application of an optimized pipeline for 454 reads identifies potential novel driver mutations in T-ALL"

2013 **ICG-Europe 2013**
Ghent, Belgium
"Identification of point mutations, expression perturbations, and gene fusions in T-cell acute lymphoblastic leukemia by RNA-seq"

Publication List

Mutation analysis of the tyrosine phosphatase PTPN2 in Hodgkin lymphoma and T-cell non-Hodgkin lymphoma.
Kleppe R, Tousseyn T, Geissinger E, **Kalender Atak Z**, Aerts S, Rosenwald, A, Wlodarska, I, Cools, J. (2011). Haematologica, 96 (11), 1723-1727.

Using cisTargetX to predict transcriptional targets and networks in Drosophila.
Potier D, **Kalender Atak Z**, Sanchez M, Herrmann C, Aerts S. (2012). Methods in Molecular Biology, 786, 291-314.

Variations in the exome of the LNCaP prostate cancer cell line.
Spans L, **Kalender Atak Z**, Van Nieuwerburgh F, Deforce D, Lerut E, Aerts S, Claessens F. (2012). The Prostate, 72 (12), art.nr. 10.1002/pros.22480, 1317-27.

High accuracy mutation detection in leukemia on a selected panel of cancer genes.
Kalender Atak Z*, De Keersmaecker K*, Gianfelici V, Geerdens E, Vandepoel R, Pauwels D, Porcu M, Lahortiga I, Brys V, Dirks W, Quentmeier H, Cloos J, Cuppens H, Uyttebroeck A, Vandenberghe P, Cools J, Aerts S. (2012). PLoS One, 7 (6), e38463.

Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia.
De Keersmaecker K*, **Kalender Atak Z***, Li N*, Vicente C, Patchett S, Girardi T, Gianfelici V, Geerdens E, Clappier E, Porcu M, Lahortiga I, Lucà R, Yan J, Hulselmans G, Vranckx H, Vandepoel R, Sweron B, Jacobs K, Mentens N, Wlodarska I, Cauwelier B, Cloos J, Soulier J, Uyttebroeck A, Bagni C, Hassan B, Vandenberghe P, Johnson A, Aerts S, Cools J. (2013). Nature genetics, 45 (2), art.nr. 10.1038/ng.2508, 186-90.

The Drosophila homologue of the amyloid precursor protein is a conserved modulator of Wnt PCP signaling.
Soldano A, Okray Z, Janovska P, Tmejová K, Reynaud E, Claeys A, Yan J, **Atak Z**, De Strooper B, Dura J, Bryja V, Hassan B. (2013). PLoS Biology, 11 (5), art.nr. 10.1371/journal.pbio.1001562, e1001562.

Identification of a novel, recurrent MBTD1-CXorf67 fusion in low-grade endometrial stromal sarcoma.
Dewaele B, Przybyl J, Quattrone A, Finalet Ferreiro J, Vanspauwen V, Geerdens E, Gianfelici V, **Kalender Atak Z**, Wozniak A, Moerman P, Sciot R, Croce S, Amant F, Vandenberghe P, Cools J, Debiec-Rychter M.(2013). International Journal of Cancer, Ahead of print, art.nr. 10.1002/ijc.28440.

Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia
Kalender Atak Z*, Gianfelici V*, Hulselmans G*, De Keersmaecker K*, Devasia AG, Geerdens E, Mentens N, Chiaretti S, Durinck K, Uyttebroeck A, Vandenberghe P, Wlodarska I, Cloos J, Foa R, Speleman F, Cools J, Aerts S. (2013). PLoS Genetics (manuscript accepted)

* equal contribution